

Excel 2007 数据挖掘完全手册

谢邦昌　朱建平　来升强　编著

系统性　详细叙述数据挖掘的一般概念、通行规范、方法技术以及软件应用等，使读者获得一个较为清晰和正确的数据挖掘观念

具体性　围绕Excel 2007的数据挖掘模块，对Excel 2007强大的表格工具详加讲解，有助于读者在工作表中完成种种复杂的数据分析任务

实用性　提供一些大型应用案例，通过详细的操作讲解和结果解释，可令读者获得实际的数据挖掘经验，从而能迅速加以应用

清华大学出版社

高效办公“职”通车

Excel 2007 数据挖掘完全手册

谢邦昌　　朱建平　来升强　编著

清华大学出版社
北　京

内 容 简 介

本书围绕 Excel 2007 的数据挖掘模块，通过大量操作示范，介绍了主流的数据挖掘方法。全书包括数据挖掘算法介绍、Excel 2007 数据挖掘模块介绍、其他分析工具介绍、数据挖掘范例 4 篇，共 26 章。除了给出有关的理论和原理阐述之外，还提供了一些大型应用案例。通过详细的操作讲解和结果分析，读者可以获得实际的数据挖掘经验，并能迅速地在自己所处的领域中加以应用。

利用 Excel 2007 的数据挖掘模块，读者无须经过专业培训，就能完成多种数据挖掘任务。本书适用于学习数据挖掘和相关课程的学生、运用 Excel 2007 进行复杂大型数据分析的职场人士及咨询公司从业人员等。

图书在版编目（CIP）数据

Excel 2007 数据挖掘完全手册/谢邦昌，朱建平，来升强编著．—北京：清华大学出版社，2008.7
（高效办公“职”通车）

ISBN 978-7-302-17474-5

I．E…　II．①谢…　②朱…　③来…　III．电子表格系统，Excel 2007－手册　IV．TP391.13-62

中国版本图书馆 CIP 数据核字（2008）第 057618 号

责任编辑：吴颖华　孙　斌
封面设计：张　岩
版式设计：牛瑞瑞
责任校对：马军令
责任印制：
出版发行：清华大学出版社　　**地　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　编**：100084
社 总 机：010-62770175　　**邮　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：
装 订 者：
经　销：全国新华书店
开　本：185×260　**印　张**：19.25　**字　数**：426 千字
版　次：2008 年 7 月第 1 版　**印　次**：2008 年 7 月第 1 次印刷
印　数：1～5000
定　价：32.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：010-62770177 转 3103　产品编号：027874-01

丛　书　序

随着计算机技术的全面迅速发展，人们将在行政办公、财务管理、会计、统计、审计等众多领域面对计算机的应用和管理。掌握计算机在这些领域的应用，一方面可以极大地提高工作效率，另一方面也可以提高业务水平。信息时代，许多行业都要求工作者有很强的计算机操作技能，做到运用自如，熟练而且深入地掌握软件的应用。而要做到这一点，必须从软件的实际应用入手。

正是在这一大背景下，我们策划了本套丛书，精选了应用领域较广泛、较常用的一些软件，如 Excel、SPSS、用友财务软件、金蝶财务软件等，旨在帮助广大办公人员、财务人员、统计分析人员、审计人员及相关专业的学生快速掌握这些软件的应用，用以解决实际工作或学习中的问题，提高自身的应用水平。

内容安排

本丛书强调软件与职业应用相结合，以实例为载体，着重介绍常用软件的操作功能和实践应用技巧。本套丛书包括：

《用友 ERP-U8 财务软件应用实务》
《金蝶 K/3 财务软件应用实务》
《SPSS 在统计分析中的应用》
《Excel 2007 在统计分析中的应用》
《Excel 2007 在会计工作中的应用》
《Excel 2007 在财务管理中的应用》
《Excel 2007 在审计分析中的应用》
《Excel 2007 函数、公式范例应用》
《Excel 2007 数据图表范例应用》
《Excel 2007 VBA 办公范例应用》
《Excel 2007 数据挖掘完全手册》

丛书特色

1．软件与职业应用相结合，实用性强。深入浅出地讲述了行政办公、财务管理、会计、统计、数据挖掘、审计各职业领域的关键知识，系统介绍了相应的软件应用方法及技巧，对实际工作有极大的帮助和指导意义。

2．内容丰富，案例典型。本套书每章都有实践案例，读者可以根据自己的情况进行取舍，直接应用于具体的工作之中。

3．结构合理，逻辑清晰。从全新的实例角度出发，按照“基本知识点讲解——实践应用——解决问题”的逻辑结构编写，全面介绍了这些软件在日常工作中的应用。符合读者的学习思路，可以使广大读者在最短的时间内学习并利用应用软件的各种强大功能，少走

弯路，迅速提升专业技能和提高工作效率。

4．光盘特色。本套丛书大部分都配有光盘，汇集了书中所用的应用软件、实例素材，及应用实例的视频，极大地方便了读者的学习。

读者定位

1．适合作为行政办公、财务管理、会计、统计、审计等领域在职工作人员提高自身业务水平的参考用书，还适合于非统计类的研究生及从事相关数据分析人员学习。

2．适合作为高校财务管理、会计、统计、审计、市场营销、电子商务、信息管理等相关专业的教材或学习用书。

3．适合作为各相关领域应用培训或职业培训的教学用书。

售后服务

如果读者在阅读图书的过程中有什么问题或需要帮助，可以登录本丛书的信息支持网站 http://www.thjd.com.cn 或通过 zzfangcn@vip.163.com（010-62788951-269）联系，也可以在 http://www.thjd.com.cn 的读者留言栏目留言，我们将尽快给您提供帮助与支持。

前　　言

目前，各行各业都开始利用计算机及相应的信息技术进行管理和决策，这使得各企事业单位生成、收集、存储和处理数据的能力大大提高。数据量与日俱增，大量复杂信息层出不穷，人们将面临着复杂数据的处理问题。Excel 是当前使用最普遍的电子表格软件，它能容易地完成图表的制作、统计、分析以及数据处理，不但功能强大，而且简单易用。最新版本的 Microsoft Office Excel 2007 支持超过 104 万笔记录的单张数据工作表，并可以同时存储 1.6 万列的数据。为能有效提升 Excel 2007 用户数据处理和分析的能力，微软公司提供了一个免费的数据挖掘模块。通过调用该模块，Excel 2007 用户可以方便快速地完成以往只有使用专业数据挖掘软件才能完成的任务。因此，我们编写了《Excel 2007 数据挖掘完全手册》这本书，其目的是使具有一定 Excel 基础的读者，能够在了解相关统计思想与方法的基础上，运用该软件对复杂数据和海量数据进行处理、分析。

本书的编写力求以统计思想为主线，以数据挖掘技术应用为目的。基本内容和特点具体体现为：第 1 篇详细叙述数据挖掘的一般概念、通行规范、方法技术以及软件应用等，使读者获得一个较为清晰和正确的数据挖掘观念。第 2 篇围绕 Excel 2007 的数据挖掘模块，通过大量操作示范，详细讲述了 Excel 2007 数据挖掘模块的九大模型的使用。这些模型包括决策树、贝叶斯概率分类、关联规则、聚类分析、时序聚类、线性回归、Logistic 回归、类神经网络和时间序列分析，基本涵盖了主要的数据挖掘技术和方法。第 3 篇介绍了 Excel 2007 的其他分析工具，结合数据挖掘技术和方法，使用改进的 Excel 表格工具，可以很方便地进行图形化的分析。第 4 篇是数据挖掘的案例分析，包括投资决策、信用评级，以及市场销售和客户细分等领域的数据挖掘模型。通过详细的操作讲解和结果解释，读者可以获得实际的数据挖掘经验，并能迅速在自己所处的领域中加以应用。

本书适合多层次多专业人士如数学、统计、经济金融、管理类等专业的本科生、专科生学习，还适合于非统计类的研究生及从事相关数据分析的人员阅读。

本书在编写及出版的过程中，得到了厦门大学经济学院计划统计系、台湾辅仁大学统计资讯学系和清华大学出版社的大力支持，在此一并表示衷心感谢！编写一本好书并不容易，尽管我们努力想奉献给读者一本满意的书，但仍有一些内容达不到读者各方面的要求。书中难免有疏漏之处，恳请读者多提宝贵意见，以便今后进一步修改与完善。

为了方便读者高效、便捷地使用本书，特免费提供本书所有实例的原始数据、源文件，请登录清华大学出版社网站（www.tup.tsinghua.edu.cn）下载。

本书的编写得到了厦门大学讲座教授基金和国家教育部“新世纪优秀人才支持计划”（Program for New Century Excellent Talents in University，NCET）的资助。

编　者

2008 年 3 月

目　　录

第 1 篇　数据挖掘算法介绍

第 1 章　数据挖掘简介 3
1.1　数据挖掘的定义 3
1.2　数据挖掘的重要性 3
1.3　数据挖掘的功能 3
1.4　数据挖掘的步骤 4
1.5　数据挖掘建模的标准 CRISP-DM 5
第 2 章　数据挖掘运用的理论和技术 7
2.1　回归分析 7
2.1.1　简单线性回归分析 7
2.1.2　多元回归分析 7
2.1.3　岭回归分析 8
2.1.4　Logistic 回归分析 9
2.2　关联规则 9
2.3　聚类分析 10
2.4　判别分析 11
2.5　类神经网络分析 12
2.6　决策树分析 13
2.7　其他分析方法 15
第 3 章　数据挖掘与相关领域的关系 17
3.1　数据挖掘与统计分析的不同 17
3.2　数据挖掘与数据仓储的关系 17
3.3　知识发现与数据挖掘的关系 18
3.4　OLAP 与数据挖掘的关系 19
3.5　数据挖掘与机器学习的关系 19
3.6　网络挖掘与数据挖掘的关系 20
第 4 章　数据挖掘商业软件产品及其应用现状 21
4.1　数据挖掘商业软件的分类 21
4.2　主要软件的介绍 21

4.3 顾客关系管理......22
4.4 数据挖掘的行业应用......23

第 2 篇 Excel 2007 数据挖掘模块介绍

第 5 章 安装与设定 Excel 2007 数据挖掘加载项......27
5.1 系统需求......27
5.2 开始安装......27
5.3 完成安装验证......30
5.4 组件设定......30
5.5 配置完成检查......35
第 6 章 Excel 2007 数据挖掘入门......37
6.1 Excel 2007 数据挖掘功能介绍......37
6.2 数据挖掘使用说明......37
6.2.1 目录查询......37
6.2.2 开始功能......38
6.2.3 视频和教学......39
6.3 数据挖掘连接配置......39
6.3.1 设定目前的连接......39
6.3.2 跟踪......41
6.4 数据准备......41
6.4.1 浏览数据......41
6.4.2 清除数据......44
6.4.3 分割数据......46
6.5 数据建模......50
6.6 精确度和验证......51
6.6.1 准确性图表......51
6.6.2 分类矩阵......52
6.6.3 利润图......53
6.7 模型用法......53
6.7.1 浏览功能......53
6.7.2 查询功能......56
6.8 模型管理......57
6.8.1 重新命名挖掘模型......57
6.8.2 删除挖掘结构......57
6.8.3 清除挖掘结构......58
6.8.4 用原始数据处理挖掘结构......58

6.8.5 用新数据处理挖掘结构......58
6.8.6 导出挖掘结构......59
6.8.7 导入挖掘结构......60

第 7 章 决策树......61

7.1 基本概念......61
7.2 决策树模块的建立......61
7.3 决策树与判别函数比较......61
7.4 计算方法......62
7.4.1 确定预测精度的标准......62
7.4.2 选择分裂（分层）技术......62
7.4.3 定义停止分裂（分层）的时间点......62
7.4.4 选择适当大小的决策树......63
7.5 Excel 2007 决策树算法......63

第 8 章 贝叶斯概率分类......71

8.1 基本概念......71
8.2 Excel 2007 贝叶斯概率分类......73

第 9 章 关联规则......84

9.1 基本概念......84
9.2 关联规则的种类......85
9.3 关联规则的算法：Apriori 算法......85
9.4 Excel 2007 关联规则......86

第 10 章 聚类分析......96

10.1 基本概念......96
10.2 层次聚类分析......96
10.3 聚类分析原理......97
10.4 Excel 2007 聚类分析......100

第 11 章 时序聚类......116

11.1 基本概念......116
11.2 相关研究和算法......116
11.3 Excel 2007 时序聚类......117

第 12 章 线性回归......126

12.1 基本概念......126
12.2 简单回归分析......127
12.3 多元回归分析......130

12.4 Excel 2007 线性回归 133

第 13 章 Logistic 回归 142

13.1 基本概念 142
13.2 logit 变换 142
13.3 Logistic 分布 143
13.4 列联表的 Logistic 回归模型 144
13.5 Excel 2007 Logistic 回归 145

第 14 章 类神经网络 161

14.1 基本概念 161
14.2 类神经网络的架构与训练算法 163
14.3 类神经网络的特性 163
14.4 类神经网络应用 163
14.5 类神经网络优缺点 164
14.6 Excel 2007 类神经网络 165

第 15 章 时间序列分析 175

15.1 基本概念 175
15.2 时间序列的成分 177
15.3 时间序列数据的图形介绍 178
15.4 利用平滑法预测 182
15.5 用趋势方程预测时间序列 186
15.6 预测含趋势与季节成分的时间序列 187
15.7 利用回归模型预测时间序列 188
15.8 其他预测模型 189
15.9 单变量时间序列预测模型 189
15.10 时间趋势预测模型 192
15.11 Excel 2007 时间序列 193

第 16 章 DMX 介绍 198

16.1 DMX 介绍 198
16.2 DMX 函数介绍 199
16.2.1 模型建立 200
16.2.2 模型训练 201
16.2.3 模型使用（预测） 201
16.2.4 其他函数语法 202
16.3 DMX 数据挖掘语法 205
16.3.1 决策树 206

16.3.2 贝叶斯概率分类......207
16.3.3 关联规则......207
16.3.4 聚类分析......208
16.3.5 时序聚类......209
16.3.6 线性回归......210
16.3.7 Logistic 回归......211
16.3.8 类神经网络......212
16.3.9 时间序列......213
16.4 DMX 应用范例......214
16.4.1 分类......215
16.4.2 估计......216
16.4.3 预测......217
16.4.4 关联分组......217
16.4.5 聚类......218

第 3 篇 其他分析工具介绍

第 17 章 分析关键影响因素......223
第 18 章 检测类别......228
第 19 章 从示例填充......231
第 20 章 预测......233
第 21 章 突出显示异常值......235
第 22 章 应用场景分析......238
22.1 目标查找......238
22.2 假设......240
第 23 章 Visio 2007 数据透视分析......243

第 4 篇 数据挖掘范例

第 24 章 上市公司投资价值分析的挖掘模型......251
24.1 研究动机与目的......251
24.2 挖掘模型的构建......251
24.3 变量筛选......252
24.4 决策树模型......253
24.5 贝叶斯概率模型......255
24.6 Logistic 回归模型......255
24.7 预测准确度比较......256

第 25 章 信用卡用户信用评测的挖掘模型......259
25.1 研究背景......259
25.2 研究动机......260
25.3 研究目的......260
25.4 Excel 2007 构建数据挖掘模型......260
25.4.1 决策树分析......260
25.4.2 聚类分析......263
25.4.3 Logistic 回归......269
第 26 章 市场营销与客户细分的挖掘模型......271
26.1 研究动机与目的......271
26.2 研究方法与限制......271
26.3 数据分析......271
26.4 挖掘建模......273
26.4.1 决策树......273
26.4.2 单纯贝叶斯分类......280
26.4.3 聚类分析......282
26.4.4 决策树......286
26.4.5 Logistic 回归......288
26.4.6 关联分析......292
26.5 结论......295



第1篇

数据挖掘算法介绍

- 数据挖掘简介
- 数据挖掘运用的理论和技术
- 数据挖掘与相关领域的关系
- 数据挖掘商业软件产品及其应用现状

第 1 章　数据挖掘简介

1.1　数据挖掘的定义

Data mining is the process of seeking interesting or valuable information in large database.

数据挖掘（data mining）是近年来数据库应用领域中相当热门的话题。数据挖掘一般是指在数据库或数据仓库中，利用各种分析方法与技术，对过去累积的大量繁杂数据进行分析、归纳与整合等工作，提取出有用的信息，例如趋势（trend）、模式（pattern）及相关性（relationship）等，并将其中有价值的信息作为决策参考提供给决策者。通俗地说，数据挖掘就是从数据中发掘信息或知识，有人称为知识发现（knowledge discovery in database，KDD），也有人称为数据考古学（data archeology）、数据模式分析（data pattern analysis）或功能相依分析（functional dependency analysis）。目前，数据挖掘已经成为数据库系统、机器学习、统计方法等多个学科相互交叉的重要领域，而在实务界，越来越多的企业开始认识到，实施数据挖掘可以为企业带来更多潜在的商业机会。

但我们对数据挖掘应有一个正确的认知：数据挖掘不是一个无所不能的魔法。数据挖掘的种种工具都是从数据中发掘出各种可能成立的“预言”，并对其潜在价值加以“估计”，但数据挖掘本身并不能在实际中查证和确认这些假设，也不能判断这些假设的实际价值。

1.2　数据挖掘的重要性

现代企业经常会搜集大量的数据，这些数据涵盖了市场、客户、供货商，及其竞争对手等重要信息，但是由于信息超载与无结构化，企业的决策者无法充分利用这些庞大的数据资源，仅能使用其中的一小部分，这可能导致决策失误，甚至出现决策错误。而借助数据挖掘技术，企业完全有能力从浩瀚的数据海洋中，挖掘出全面而又有价值的信息和知识，并作为决策支持之用，进而形成企业独有的竞争优势。

1.3　数据挖掘的功能

一般而言，数据挖掘包括下列五项功能，这些功能大多为成熟的计量和统计分析方法。

1．分类（classification）

按照分析个体的属性状态分别加以区分，并建立类组（class）。例如，将信用申请者的风险等级分为高风险、中风险和低风险三类。使用的方法有决策树（decision tree）、判别分

析（discriminant analysis）、类神经网络（artificial neural network），以及记忆基础推理（memory-based reasoning）等。

2．估计（estimation）

根据已有的数值型变量和相关的分类变量，以获得某一属性的估计值或预测值。例如，根据信用卡申请者的教育程度和从事职业来设定其信用额度。使用的方法有相关分析、Logistic 回归及类神经网络等。

3．预测（prediction）

根据个体属性的已有观测值来估计该个体在某一属性上的预测值。例如，由顾客过去刷卡消费额预测其未来的刷卡消费额。使用的方法有回归分析、时间序列分析及类神经网络等。

4．关联分组（affinity grouping）

从所有对象决定哪些相关对象应该放在一起。例如，超市中相关的洗漱用品（牙刷、牙膏、牙线）放在同一货架上。在客户营销系统上，这类分析可以用来发现潜在的交叉销售（cross-selling）商品聚类，进而设计出有价值的组合商品集合。

5．同质分组（clustering）

将异质总体分成为同质性的类别（clusters），即聚类。其目的是识别出总体中所包含的混合类别的组间差异，并根据每个类别的特征对所有个体进行归类。同质分组相当于营销术语中的细分（segmentation）。应该注意的是：聚类分析根据数据自动产生各个类别，事先是不知道或无须知道总体中潜在的类别信息。使用的方法有 k-means 等动态聚类法及 agglomeration 等层次聚类法。

1.4　数据挖掘的步骤

数据挖掘的步骤会随不同领域的应用而有所变化，每一种数据挖掘技术也会有各自的特性和使用步骤，针对不同问题和需求所制定的数据挖掘过程也会存在差异。此外，数据的完整程度、专业人员支持的程度等都会对建立数据挖掘过程有所影响（蔡维欣，2003）。这些因素造成了数据挖掘在各不同领域中的运用、规划，以及流程的差异性，即使同一产业，也会因为分析技术和专业知识的涉入程度不同而不同，因此对于数据挖掘过程的系统化、标准化就显得格外重要。如此一来，不仅可以较容易地跨领域应用，也可以结合不同的专业知识，发挥数据挖掘的真正精神。

数据挖掘完整的步骤如下：

① 理解数据和数据的来源（understanding）。

② 获取相关知识与技术（acquisition）。

③ 整合与检查数据（integration and checking）。

④ 去除错误或不一致的数据（data cleaning）。

⑤ 建立模型和假设（model and hypothesis development）。
⑥ 实际数据挖掘工作（data mining）。
⑦ 测试和验证挖掘结果（testing and verfication）。
⑧ 解释和应用（interpretation and use）。

由上述步骤可看出，数据挖掘牵涉了大量的准备工作与规划工作，事实上许多专家都认为整套数据挖掘的过程中，有 80%的时间和精力是花费在数据预处理阶段，其中包括数据的净化、数据格式转换、变量整合，以及数据表的链接。可见，在进行数据挖掘技术的分析之前，还有许多准备工作要完成。

1.5　数据挖掘建模的标准 CRISP-DM

CRISP-DM 是 Cross-Industry Standard Process for Data Mining 的简称，中文翻译为“数据挖掘的跨行业标准过程”。CRISP-DM 是由欧洲几家在数据挖掘应用上有经验的公司共同筹划组织的一个特别小组所提出的。该组织的成员包括数据仓储供货商 NCR、德国汽车航天公司 Daimler-Chrysler、统计分析软件供货商 SPSS 和荷兰的银行保险公司 OHRA，除了 NCR 与 SPSS 等是专注于数据挖掘软件开发的成员之外，也有其他众多厂商参与实验，通过实际操作过程，整体规划设计，并在 2000 年推出了 CRISP-DM 1.0 模型，把数据挖掘过程中必要的步骤都加以标准化。CRISP-DM 模型强调完整的数据挖掘过程，不能只针对数据整理、数据显示、数据分析以及构建模型，而应该将对企业的需求问题的理解，以及后期对模型的评价与模型的延伸应用都纳入到数据挖掘过程中。因此，CRISP-DM 从方法学的角度强调了实施数据挖掘项目的方法和步骤，同时独立于每种具体数据挖掘算法和数据挖掘系统。

CRISP-DM 分为六个阶段（phase）和四个层次（level），分别简介如下。

六个阶段如下。

1．定义商业问题（business understanding）

本阶段的主要工作是要针对企业问题以及企业需求进行了解确认，针对不同的需求做深入的了解，将其转换成数据挖掘的问题，并拟定初步构想。在此阶段中，需要与企业各层次进行讨论，只有对要解决的问题有了非常清楚而全面的了解，才能正确地针对问题拟定分析过程。

2．数据理解（data understanding）

此阶段包括建立数据库与分析数据。在这个阶段必须先收集数据，了解数据的含义与特性，并过滤出所有可能有用的数据，然后进行数据整理并评估数据的质量，必要时再将分属不同数据库的数据加以合并或整合。数据库建立完成后再进行数据分析，并找出影响最大的数据，进而判断是否有必要进一步收集更为详细的数据。

3．数据预处理（data preparation）

此阶段和数据理解阶段为数据准备阶段的核心，这是建立模型前的最后一步数据准备

工作。数据预处理任务很可能要反复执行多次，并且没有任何既定的顺序，其目的是把各种不同来源的数据加以清理、整理和归并，以适合数据挖掘技术的使用。

4．建立模型（modeling）

此阶段对预处理过的数据应用各种数据挖掘技术，建立分析模型，发现企业问题的根源。面对同一个问题，会有多种可供使用的分析技术，但是每种技术对数据都有不尽相同的要求，因此需要回到数据预处理阶段，重新转换数据为符合要求的格式。

5．评价和解释（evaluation and explanation）

从数据分析的观点看，在开始进入这个阶段时已经建立了看似是高质量的模型，但在实际应用中，随着应用数据的不同，模型的准确率肯定会变化。这一阶段的主要任务是对于挖掘结果加以评价和解释。一个值得注意的问题是：是否有某些重要的商业问题还没有充分地考虑，以至于使模型的预测精度发生了显著的变化。

6．实施（deployment）

一般而言，完成模型创建并不意味着项目结束。模型建立并经验证后，有两种主要的使用方法。第一种是提供给分析人员做参考，由分析人员通过查看和分析这个模型后提出行动方案建议；另一种是把此模型应用到不同的数据集上。此外，在应用了模型后，当然还要不断监控它的效果。

四个层次（level）分别为 phase、generic task、specialized task 和 process instance。每个 phase 由若干 generic task 组成，每个 generic task 又实施若干 specialized task，每个 specialized task 由若干 process instance 来完成。其中，上两层独立于具体数据挖掘方法，即是一般数据挖掘项目均需实施的步骤（What to do?），这两层的任务将结合具体数据挖掘项目的“上下文”（context）映射到下两层的具体任务和过程。所谓项目的“上下文”是指项目开发中密切相关、需要综合考虑的一些关键问题，如应用领域、数据挖掘问题类型、技术难点、工具及其提供的技术等。

第 2 章　数据挖掘运用的理论和技术

2.1　回 归 分 析

回归分析主要用于了解自变量与因变量间的数量关系。其目的是获得变量间相关性的数量描述，通过控制自变量来影响因变量，达到所谓“以价制量”的效果，也可以利用已知变量对未知变量做预测。当然，选取自变量时，必须注意所选出的自变量与因变量是否存在着因果关系。

2.1.1　简单线性回归分析

最简单的回归，只包括一个因变量 Y 与一个自变量 X，同时希望它们之间的关系是直线：

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \ , \ i = 1, 2, \cdots, n$$

其中：

Y 为因变量（dependent variable; response variable）。

X 为自变量（independent variable）。

ε 为误差项（error term）。

满足这样关系的模型，称为线性模型（linear model），模型中的参数（regression parameters）叫做回归系数（regression coefficient）。

2.1.2　多元回归分析

实际上，影响因变量 Y 的自变量 X_i 往往不只一个，而有 k 个，例如影响小麦产量的因素有雨量 X_1、气温 X_2、湿度 X_3、土壤肥力 X_4 等独立的变量。又如影响人们体重的因素有食物摄取量 X_1、运动量 X_2 及睡眠时间 X_3 三个自变量。一个因变量与多个自变量间的关系，可表示为：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \ , \ i = 1, 2, \cdots, n$$

该式中，各自变量都是一次幂式，称为多元线性回归模型，其中参数 β_0 为截距，参数 $\beta_1, \beta_2, \cdots, \beta_k$ 为回归系数。

2.1.3 岭回归分析

当自变量间存在多重共线性关系时，这些自变量就不适合放入同一模型。如果自变量间存在高度多重共线性，则回归系数的方差变大，使得一个或多个自变量因为无法通过参数的显著性检验而被舍弃，从而建立一个没有效率的回归模型。所以在建模前，需要对自变量间的多重共线性进行检查，以避开这一问题。最直接的方法是同一模型中避免选取有高度相关性的自变量；另一种方法是利用统计方法，如利用岭回归来降低回归系数估计值的方差。

多重共线性是指自变量间有比较显著的相关性。假设有 m 个自变量被考虑放入同一个回归模型中，如果利用简单的相关系数只能发现两个自变量间的相关程度，不能发现多变量之间的相关性。参照线性回归的思路，可以利用某一自变量与其他 $m-1$ 个自变量间多元回归的判定系数的大小来判断多重共线性的强烈程度。若第 i 个自变量与其他 $m-1$ 个自变量的回归方程为：

$$\hat{x}_i = S_i + t_1x_1 + \cdots + t_{i\text{-}1}x_{i\text{-}1} + t_{i+1}x_{i+1} + \cdots + t_mx_m, i = 1, 2, \cdots, m$$

其中：

S_i 为第 i 个多元回归模型的截距项；

t_m 为第 m 个多元回归模型的回归系数。

此模型得到的回归判定系数为：

$$R_i^2 = \frac{\text{SSR}_i}{\text{SST}_i},\ i = 1, 2, \cdots, m$$

可以通过计算方差膨胀因子（Variance Inflation Factor，VIF）来表示多重共线性的指数，其计算公式为：

$$\text{VIF}_i = \frac{1}{1 - R_i^2},\ i = 1, 2, \cdots, m$$

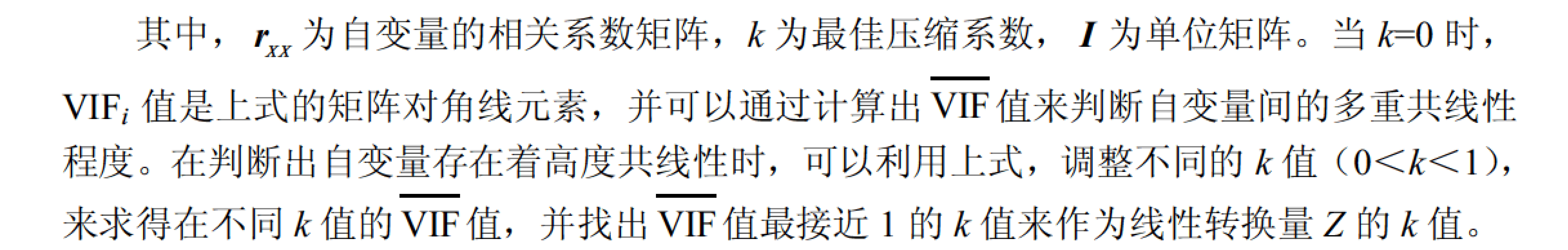
当 $R_i^2 = 0$ 时，表示第 i 个自变量与其他 $m-1$ 个自变量不相关，则 $\text{VIF}_i = 1$；而当 $R_i^2 \to 1$，表示第 i 个自变量与其他 $m-1$ 个自变量趋近于完全相关，则 $\text{VIF}_i \to \infty$，可见 VIF_i 具有测度多重共线性的能力。m 个自变量可以计算出 m 个 $\overline{\text{VIF}}$ 值，其中若是最大的 $\overline{\text{VIF}}$ 值超过 10（表明至少某个判定系数大于 0.9），则认为自变量存在着高度的多重共线性。当自变量数目过多时，可以对 m 个 $\overline{\text{VIF}}$ 值求取平均数：

$$\overline{\text{VIF}} = \frac{1}{m}\sum_{i=1}^{m} \text{VIF}_i$$

若 $\overline{\text{VIF}}$ 明显大于 1，则认为多重共线性存在。

一般地，$\overline{\text{VIF}}$ 值的计算可以利用自变量的相关系数矩阵来求得：

$$(\boldsymbol{r}_{xx} + k\boldsymbol{I})^{-1}\boldsymbol{r}_{xx}(\boldsymbol{r}_{xx} + k\boldsymbol{I})^{-1}$$



其中，r_{xx} 为自变量的相关系数矩阵，k 为最佳压缩系数，I 为单位矩阵。当 k=0 时，VIF_i 值是上式的矩阵对角线元素，并可以通过计算出 $\overline{VIF}$ 值来判断自变量间的多重共线性程度。在判断出自变量存在着高度共线性时，可以利用上式，调整不同的 k 值（0<k<1），来求得在不同 k 值的 $\overline{VIF}$ 值，并找出 $\overline{VIF}$ 值最接近 1 的 k 值来作为线性转换量 Z 的 k 值。

2.1.4　Logistic 回归分析

回归分析是利用一系列的数值型变量来预测另一个数值型变量，但无法对仅仅具有若干状态的定性变量进行预测。定性变量的分析，需要使用 Logistic 回归分析。Logistic 回归可以分析一大类的问题，例如讨论定性变量和数值变量对同一个类别变量的影响和关系；它们之间的独立性；在不独立时具有什么形式的数量关系。当因变量是一个 0/1 变量时（只取 0 和 1 两种值），如果定义 $y = 1$ 的概率 $p = \Pr\{y = 1\}$ 为要研究的对象，将影响 y 变动的因素定义为自变量，记为 $x_1,\cdots,x_k$，这其中既有定性变量，也有数值变量。线性 Logistic 回归假设自变量和因变量之间存在以下数量关系：

$$\ln\left(\frac{p}{1-p}\right) = a_0 + a_1x_1 + \cdots + a_kx_k$$

即 $\ln\left[Ey/(1-Ey)\right]$ 是关于 $x_1,\cdots,x_k$ 的线性函数。而若等式左边是一个非线性函数 $g(x_1,\cdots,x_k;c)$，其中 c 表示可能包含的参数向量，则相应模型称为非线性 Logistic 回归模型。

2.2　关 联 规 则

关联规则用于发现数据中变量间的关系。随着数据不断地收集和储存，从大量商业交易记录中会发现有趣的关联规则，有助于许多商业决策的制定，如商品组合设计和交叉销售等。

关联规则中最典型的一个应用就是购物篮分析。该方法通过记录顾客放入其购物篮中不同商品的条形码，分析顾客的购买特性。了解某些商品组合被顾客同时购买的概率高低，通过此关联的发现，可以协助零售商拟定产品组合营销策略。例如，一次超市购物中，某位顾客如果已经购买了牛奶，则其同时具有购买面包的可能性。通过帮助零售商有选择地规划商品的摆设地点和促销组合，由此引导销售，提高商品组合的销售量。

关联规则最早由 Agrawal 提出，例如两个商品项目集 X，Y 可能同时被购买，那么可以建立规则 $X \Rightarrow Y$，并采用该规则中所包含项目的联合概率来测度这一规则的发生频率高低。关联规则中有两个重要的参数：支持度（support）和可信度（confidence）。其中支持度是指 X 与 Y 同时出现在 D 交易记录数据集的次数，除以 D 中交易记录的次数的值；以概率的观点来看，支持度就是同时发生 X 与 Y 事件的联合概率。可信度是指 X 与 Y 同时出现在 D 交易总集合的次数，除以 X 项集在 D 交易总集合出现的次数的值；以概率的观点来看，可信度就是在 X 事件发生的情况下，Y 事件发生的条件概率。

2.3 聚类分析

聚类分析是一种动态分类的方法，可以把相似的事物归入合适的类别，使同类中的事物尽可能地相似（组内同质性），而类与类之间保持显著的差异（组间异质性）。例如，根据描述客户相似或差异性的指标，将客户群体分割成若干具有不同特点的类别，进而达到市场分割的目的。

在聚类分析中，所有客户所属分类是事前未知的，客户群体中存在的类别数也是未知的。为得到合理的分类，必须使用适当的指标来定量地描述研究对象间的同质性。常用的指标为“距离”和“相似系数”。假定研究个体都用“点”来表示，在聚类分析中，一般是将“距离”较近的点或“相似系数”较大的点归为同一类，将“距离”较大或“相似系数”较小的点归为不同的类别。当然，聚类分析也可以用于分析指标间的相似性，这就相当于调换个体和指标，将原指标视为个体，而将原个体视为指标。

若用 X 与 Y 表示 s 维空间中 n 个个体中的任意两个点，如果是对变量聚类，X 和 Y 分别表示 k 个变量中的任意两个，其变量维数就是样本量 n。如果是对样本做聚类，则 X 和 Y 分别表示两个个体，维数 s 就是聚类变量的个数 k。

常用的距离指标为欧氏距离（euclidean distance），其公式如下：

$$D(X,Y)=\sqrt{\sum_i (X_i-Y_i)^2},\ i=1,2,\cdots,s$$

常用的相似系数指标为余弦系数和皮尔森相关系数。

余弦系数（cosine）的公式如下：

$$S(X,Y)=(\sum_i X_iY_i)/\sqrt{(\sum X_i^2)(\sum Y_i^2)},\ i=1,2,\cdots,s$$

皮尔森相关系数（pearson correlation）的公式如下：

$$S(X,Y)=\sum_i Z_{xi}Z_{yi}/(s-1),\ i=1,2,\cdots,s$$

其中 Z_{xi} 和 Z_{yi} 表示 X 和 Y 的标准正态得分。

常用聚类分析方法分为两大类：层次聚类法（hierarchical clustering）和非层次聚类法（non-hierarchical clustering）。层次聚类法又称系统聚类法，其聚类过程可用所谓的层次结构或树状结构来描绘，具体又分为积聚法（agglomerative clustering）和分割法（divisive clustering）两种。积聚法是先把所有的个体分别作为一类，将各组组间距离最小或相似系数最大的组合并成新的一组，在聚类准则下将所有的组归并。然后对归并后所形成的新组，再次计算其组间间距或相似系数，并将各组的组间距离最小或相似系数最大的合并，依此持续合并，直到所有的个体都被归入同一组为止。而分割法正好相反，先将所有的数据看成一个群，然后在一定的分割准则下，对该整体进行分割，使每一组中的个体尽可能远离另外一组。然后分别对每组再继续分割，直到每一个体仅包含单一个体为止。

最常用的积聚法是连接法（linkage method），根据事先定义的组与组之间的距离的计

算标准，将各组逐步合并。由于聚类间距离的定义不同，又可以分为四种：

- ❑ 单一连接法（single linkage）：也称最短距离法或最近紧邻连接法，两个类之间的距离定义为分别来自两群中的个体间的最短距离，并依此类间距离选择最靠近的组来合并。
- ❑ 完全连接法（complete linkage）：也称最长距离法或最远紧邻连接法，两个组间的距离定义为分别来自两组中的个体间的最长距离，并依此类间距离选择最靠近的组来合并。
- ❑ 平均连接法（average linkage）：也称 Ward 法（Ward's procedure），其分类标准与方差分析类似。即在分组的过程中，使组内个体间的离差平方和尽可能小，而组间的离差平方和尽可能大。
- ❑ 重心法（centroid method）：两个组之间的距离定义为两组重心之间的距离，然后与连接法类似，将各个类别逐步合并。

非层次聚类法，也称为逐步聚类法、k-means 聚类法或快速聚类法，该类型的聚类法又可以分为序列阈值法（sequential threshold method）、平行阈值法（parallel threshold method）以及最佳分离法（optimizing partitioning method），其中序列阈值法事先规定一个阈值，选取一个中心点，将与该中心点的距离在阈值内的所有点都归入同一组，然后再选取一个中心，对还没有归类的点重复该过程，直到所有点都归入某一组为止。平行阈值法与序列阈值法类似，所不同的只是所有的聚类中心是同时选取的，将阈值范围内的点归到离中心最近的那一组。最佳分离法则是允许重新分配已归类的点到其他类别内，以使总体的分组标准达到优化。分组标准需要事先确定，例如取单一连接法、完全连接法或平均连接法等。

2.4 判 别 分 析

数据挖掘中的分类功能是指：在已知现有的分类下，如何建立一套判别标准，并对新样本进行分类。例如，根据消费者的一些背景数据，可以判定哪些消费者更可能是忠诚客户，判别忠诚客户与非忠诚客户的基本特征和分析特征，并可以区分哪些心理特征或生活方式特征可以作为判别或区分客户类型的标准。这些问题的性质都是相同的：即根据从个体所测定或观察到的一些指标来判断个体属于哪种类型，并对此作出区分。

判别分析就是研究判断个体所属类型的一种多元统计方法。具体地说，判别分析中的因变量或判别准则是类别变量，而自变量或预测变量基本上是等距变量。分析的过程就是建立自变量的线性组合，使之能最佳地区分出因变量的各个类别。例如，若因变量为某种产品的价格敏感型客户和非敏感型客户，而自变量为对一组消费观念的态度得分的李克特五分量表，而在判别分析中可进行的主要有：

- ❑ 建立判别函数，即找到能最恰当地区分因变量类别的自变量的线性组合，或确定事后概率，即计算每个个体落入各类别的概率。
- ❑ 检验各类别在预测变量方面是否存在显著的差异。
- ❑ 确定哪些预测变量是区分类别差异的重要变量。

- 根据预测变量的值对个体进行分类。
- 对分类的准确程度进行评估。

判别分析模型用一个或几个判别函数来表示，在有两个类别的情况只需一个判别函数。最简单也是比较常用的判别函数为线性函数：

$$D_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \cdots + b_kX_{ki}$$

其中：

D 为判别得分，D_i表示对应于第 i 个个体的得分；

b 为判别系数或权重，b_i表示对应于第 i 个自变量或预测变量的系数；

X 为自变量或预测变量，X_{ik}表示对应于第 i 个个体在第 k 个自变量上的取值。

根据所收集样本的数据，可以计算出一个判别的临界值 D_c，作为判定某个个体归属到哪一个类别的基准。在判别分析中有一个基本的假设：每一个类别都是取自一个多元正态总体的样本，而且所有正态总体的协方差矩阵或相关系数矩阵都假定是相同的。在数据挖掘的实际应用中，常用的办法是将原始数据经过抽样后，抽出两部分，其中一部分作为分析样本（训练样本），对其进行分析并建立判别函数，再利用另外一个样本（验证样本），来评估判别函数的效果。

2.5　类神经网络分析

类神经网络的相关研究与其应用范围在近年来发展极为迅速，其应用领域包括工业工程、商业与金融、社会科学及科学技术等。其最大优点除可应用于构建非线性模型外，无须像传统统计方法那样在构建模型之前需要验证假设是否成立。类神经网络的原始想法与基本结构都和神经生物学中的神经元构造相似。根据 Freeman（1992）的定义，类神经网络是模仿生物神经网络的信息处理系统，通过使用大量简单连接的人工神经元来模仿生物神经网络的能力。而在一个网络模型中，一个人工神经元将从外界环境或其他人工神经元取得信息，根据信息的相对重要程度给予不同的权重，加总后再由人工神经元中的数学函数进行转换，并输出其结果到外界环境或其他人工神经元当中。其运作概念如图 2-1 所示。

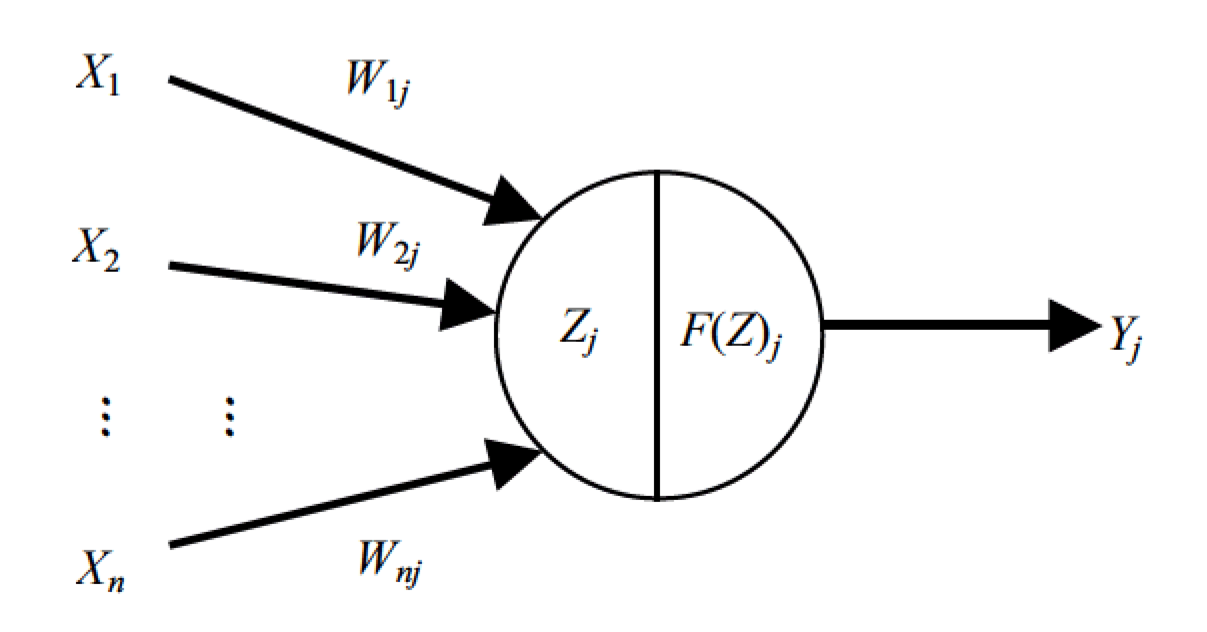


图 2-1　类神经网络原理

- X_n 为神经元的输入（input）。

- W_{nj} 为键结值（weights），类神经网络的训练就是在调整键结值，使其变大或减小，通常是通过随机的方式产生一个介于+1～-1 之间的初始值。键结值可视为一种加权效果，其值越大，则代表连结的神经元更容易被激发，对类神经网络的影响也更大；反之，则代表对类神经网络并无太大影响，而太小的键结值通常可以舍去，以节省计算机计算的时间与空间。
- Z_j 为加总单元（summation），此部分是将每一个输入值与键结值相乘后做一加总的动作。
- $F(Z_j)$ 为激活函数（activation function），通常是非线性函数，有多种不同的函数类型，其目的是将 Z 的值做映射得到所需要的输出。
- Y_j 为输出（output），也就是最终所需要的结果。

将上述的神经元组合起来就成为一个类神经网络。到目前为止，许多学者针对不同的研究问题，提出了许多类神经网络模型，各种类神经网络的算法并不相同。常见的网络有：反向传递网络、霍普菲尔网络和半径式函数网络，这些类神经网络并非适用所有的问题，必须针对欲解决问题的不同选择适当的类神经网络。

类神经网络必须通过反复训练的方式，才能获得比较好的估计参数。因此在类神经网络的学习过程中，必须提供一个训练样本，训练样本来自于实际系统输入与输出数据或以往的经验。类神经网络的工作性能与训练样本有直接的关系，若训练样本不正确、太少或太相似，类神经网络的工作适应能力与预测能力将大打折扣。换句话说，训练样本相当于类神经网络的“老师”，因此训练样本越多、越真实、差异性越大，类神经网络的能力就越强。

训练类神经网络的目的，就是让类神经网络的输出尽可能接近目标值，即相同的输入进入到系统与类神经网络，得到的输出值要尽可能相同。类神经网络在刚开始训练的时候，其输出是凌乱的。伴随着训练次数的增加，类神经网络的键结值会逐渐被训练数据调整，使得类神经网络的输出结果与目标值的误差越来越小。

学习率在类神经网络的训练过程中是一个非常重要的参数，学习率影响着类神经网络收敛的速度，若学习率选择较大则类神经网络收敛的速度较快，但其适应性将会降低；反之，较小的学习率会使得类神经网络的收敛速度变慢，但却更加稳定。选择太大或太小的学习率对类神经网络的训练都有不良的影响。

在类神经网络的训练过程中，虽然类神经网络的输出已经与所要求的数值接近，但对于非训练样本的输入，并不知道会得到何种输出。因此必须使用另一组类神经网络从未见过的样本，对经过训练的类神经网络测试其结果是否与所要求的值接近，这种用途的样本称为测试样本。如果测试样本与训练样本的预测效果差异过大，表示类神经网络模型缺乏适应性，必须重新进行训练，或者调整模型结构。

2.6　决策树分析

决策树是进行分类和预测的常用方法，采用树枝状来展现数据受各变量影响情形的预测模型，能利用树形图的分割自动确认和评估分割。由树形图可获取个体中的最佳聚类，

再通过收益图，可方便地在不同判别变量的分割点之间进行成本和效益的比较，并找出最佳获利的分割临界点。决策树和类神经网络不同，决策树中产生的规则可以用文字或数字给出明确的表达。

在数据分析中，常会遇到变量间不仅存在相关性，而且存在交互的影响关系。当两个或两个以上变量间存在交互影响时，某一变量数值改变所引起的反应，将受制于其他变量数值的大小。在商业上，研究人员通常不能确定哪几个变量间存在交互影响关系，如果预测变量数目众多，模型就会变得庞大复杂，加上预测变量间的交互影响关系可能为乘法关系，也可能为非乘法关系，大大增加了建模的难度。这时使用决策树分析，就可以较好地发现变量之间的交互关系。常用的决策树方法有 CHAID（chi squared automatic interaction detection）。CHAID 只能处理类别变量，如果是连续变量必须采用离散化处理，先转换数据成为类别变量，才可以使用。CHAID 的基本分析过程如下（黄登源，2003）：

① 针对每一变量计算其所有可能把原样本分成两个部分的分割方式，以找出一个最佳分割方式。所谓“最佳”是指数据经过分割后，准则变量的组间差异为最大。假设 Y 代表准则变量，样本数为 n，如果对预测变量一无所知，则 $\overline{Y}$ 可为最佳估计值，而 Y 的误差平方和为：

$$\sum (Y_i-\overline{Y})^2=\sum Y_i^2-n\overline{Y}^2$$

假设将原样本分割成两组，各组所含样本数为 n_1 和 n_2，各组准则变量的平均数分别为 $\overline{Y}_1$ 和 $\overline{Y}_2$，其误差平方和为：

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y})^2=\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2+\sum_{i=1}^{2}n_i\overline{Y}_i^2-n\overline{Y}^2$$

$$(\sum_{i=1}^{2}\sum_{j=1}^{n_i}Y_{ij}^2-n\overline{Y}^2)-\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_i)^2=n_1\overline{Y}_1^2+n_2\overline{Y}_2^2-n\overline{Y}^2$$

若通过分割，则误差平方和将会降低，若此值为正，表示 $n_1\overline{Y}_1^2+n_2\overline{Y}_2^2$ 大于 $n\overline{Y}^2$。经过分割成两组后，其同质性已经提高，即分割后减少的误差平方和为最大，也就是 $\max\left\{n_1\overline{Y}_1^2+n_2\overline{Y}_2^2-n\overline{Y}^2\right\}$。

② 比较各预测变量在“最佳分割方式”下的组间方差，然后找出一个组间方差最大的变量，即为最佳的预测变量。

③ 通过最佳预测变量得到的最佳分割方式把原始数据分割成两组。

④ 将分割后两组样本的每一组都作为新样本，并分别对每一组重复上述步骤，进一步进行分割。

⑤ 重复上述步骤，直到所有的个体都被分割成单独一组为止。

实际应用中，通常事先确定一些控制参数或限制条件，适时停止分割过程。例如分割后所减少的准则变量误差平方和必须超过所确定的水平时才可以继续将样本分割；或当任一组样本的误差平方和必须大于所确定的水平，才可以继续进一步分割；研究人员也可以针对原始样本分割的组数加以给定，或依据每组中的样本有多少笔数据等限制条件来设定。

2.7 其他分析方法

一种可以应用于数值型变量的决策树分析是分类回归树，即 CART（classification and regreesion tree）。该算法由 Brieman（1984）提出，采用来自经济学的分散度量法。CART 借助一个输入变量的函数，以递归的方式不断地将不同属性的个体分开，最后同属性的数据将会被归入同一个多维矩形区域，在每个区域中分别利用回归的方式进行拟合。需注意的是，CART 的层数不宜过多或过少。如果太少，即表示分割过程太早结束，所构建的模型未必产生良好的分类规则；相反，过多的层次则表示其分割过多，所产生的规则的分类能力并不理想。各种分析方法如表 2-1 所示。

表 2-1 各类分析方法整理

类别	模型	摘要	
分类技术	分类	❑ 根据一些变量的数值做计算，再依照结果分类 ❑ 用一些根据历史经验或已经分类好的数据来研究它们的特征，然后再根据这些特征对其他未经分类或新的数据做预测	
	聚类	❑ 将数据分类，其目的在于将类间的差异找出来，同时也将类别内成员的相似性找出来 ❑ 与分类不同，分析前并不知道会以何种方式或依据来分类，所以必须要配合专业领域知识来解释这些分类的意义	
	理论技术	传统技术（统计分析）	❑ 因子分析（factor analysis）——约简变量 ❑ 判别分析（discriminant analysis）——分类 ❑ 聚类分析（cluster analysis）——识别类组
		改良技术	❑ 决策树（decision tree）——用树型结构展现数据在受各变量影响的情况下得到的预测模型，根据对目标变量的状态不同而建立分类规则 ❑ 多用于客户资料的分析 ❑ 常用的分类方法为 CART 和 CHAID 两种
估计预测类	回归	❑ 使用一系列的数值来预测一个连续数值的可能值 ❑ 可利用 Logistic 回归来预测类别变量	
	时间序列	❑ 用现有的数值来预测未来的数值 ❑ 与回归不同，时间序列所分析的数值都与时间有关	
	理论技术	传统技术（统计分析）	❑ 回归——连续变量 ❑ Logistic 回归——类别变量 ❑ 时间序列——与时间相关的变量
		改良技术	❑ 类神经网络——模仿人脑思考结构的数据分析模型，根据输入变量与目标变量进行自主学习，并根据学习得到的知识不断调整参数来建立数据模型

续表

类别	模型	摘要	
估计预测类	理论技术	改良技术	❑ 传统回归分析：优点是在进行分析时无须限定模型，特别当变量间存在交互效应时可自动检测出来 ❑ 类神经网络多用于数据属于高度非线性且变量中具有相当程度的交互效应的情形
序列规则类	关联规则	❑ 找出在某一组事务中会同时出现的一些事务组合，例如，如果 *A* 是某一事件的一种选择，则 *B* 出现在该事件中的概率是多少	
	序列分析	❑ 序列分析与关联规则不同的是，序列分析事件的相关以时间因素来做分割	
	理论技术	传统技术（统计分析）	缺　乏
		改良技术	❑ 规则归纳法——由一连串的“如果……则……（if…then）”的逻辑规则对数据进行细分，在实际运用时，如何界定规则的有效性是最大的问题，通常需要先将数据中发生次数太少的样本剔除，以避免产生无意义的规则

第 3 章　数据挖掘与相关领域的关系

3.1　数据挖掘与统计分析的不同

硬要区分数据挖掘和统计学的差异其实是没有太大意义的，数据挖掘有相当大的部分是源于统计学科中的多元统计分析。但是为什么数据挖掘的出现会引起各领域的广泛注意呢？主要原因是相对于统计分析而言，数据挖掘有下列几个特性：

- 处理大型实际数据更有优势，且无须太专业的统计背景去使用数据挖掘工具。
- 数据分析趋势为从大型数据库抓取所需数据并使用专业计算机分析软件，数据挖掘的工具更符合企业需求。
- 就理论的基础点来看，数据挖掘和统计分析有应用上的差别，毕竟数据挖掘的目的是方便企业用户使用而非给统计学家检验用的。

3.2　数据挖掘与数据仓储的关系

若将 data warehousing（数据仓储）比喻为矿坑，数据挖掘就是深入矿坑挖掘的工作。毕竟数据挖掘不是一种无中生有的魔术，也不是点石成金的炼金术，若没有丰富完整的数据，是很难期待数据挖掘能挖掘出什么有意义的信息的。

从数据仓储挖掘有用的数据，则是数据挖掘的研究重点，但两者的本质与过程是两码事。换句话说，数据仓储应先行建立完成，数据挖掘才能有效进行，因为数据仓储本身所含数据应是正确的（不会有错误的数据掺杂其中）、完整的，而且是经过整合的。因此，两者的关系可以简单表示为“数据挖掘是从巨大数据仓储中找出有用信息的一种过程与技术”。

数据仓储和数据库虽然同是数据存储的手段，但两者相差甚远，数据仓储与数据库的比较如表 3-1、表 3-2 所示。

表 3-1　数据仓储和数据库的结构比较

结　构	数 据 仓 储	传统数据库
主要目的	信息取得与分析	支持每日交易数据
架构	关系型数据库管理系统	关系型数据库管理系统
数据模型	星型纲要（star schema）	正规划表格（normalized relations）
查询方式	通过 OLAP 或 MOLAP 接口	SQL
数据形式	分析性数据	交易性数据
数据储存状况	历史性、描述性数据	经常改变的、实时性的数据

表 3-2　数据仓储和数据库的特性比较

特　　性	数　据　库	数 据 仓 储
数据的时间性	当时的运算数据	经过处理的历史数据
数据库的规划方式	由下往上（bottom-up）	由上往下（top-down）
数据库的纲要设计	个体—关系模型配合正规化	星型纲要（star schema）
数据	无重复储存	大量重复储存，并预先加总
数据维护者	数据库管理师（DBA）	数据品管师（DQM）
异动的频率	经常异动（故称 OLTP）	少有异动，大多为查询
异动的数据数量	平时均有大量的异动处理	定期大量加载并聚合加总
效能要求	须能承受大量的更新要求	查询速度足够快
查询的频率	少量需求	大量需求（故称 OLAP）
查询的范围	较狭隘	相当宽广
查询的复杂度	较单纯	相当复杂
所内含的数据量	MB 级	GB 级
内含数据的错误率	可以容忍错误与缺项存在	极少错误与数据缺项
数据的精细度	存放一笔交易的详细数据	存放大量加总过的数据
整合性	依功能分数据库，未整合	整个组织的数据完全整合
主题性	依功能导向区分数据库	依主题导向
随时间变动的特性	很少会依时间流逝增加内容	依时间的流逝而增加其内容
暂存性	只保留目前最新的数据	完整保留所有历史数据
适合构建的系统	关系型数据库管理系统	多维数据库管理系统

3.3　知识发现与数据挖掘的关系

根据 Fayyad 等人（1996）对知识发现（knowledge discovery in database）的定义——它是指从数据中提取有效、全新、潜在有用、最终可被理解的模式的一个非细琐（nontrivial）的流程，其最终目标是了解数据的模式（patterns）。知识发现的主要步骤如图 3-1 所示。

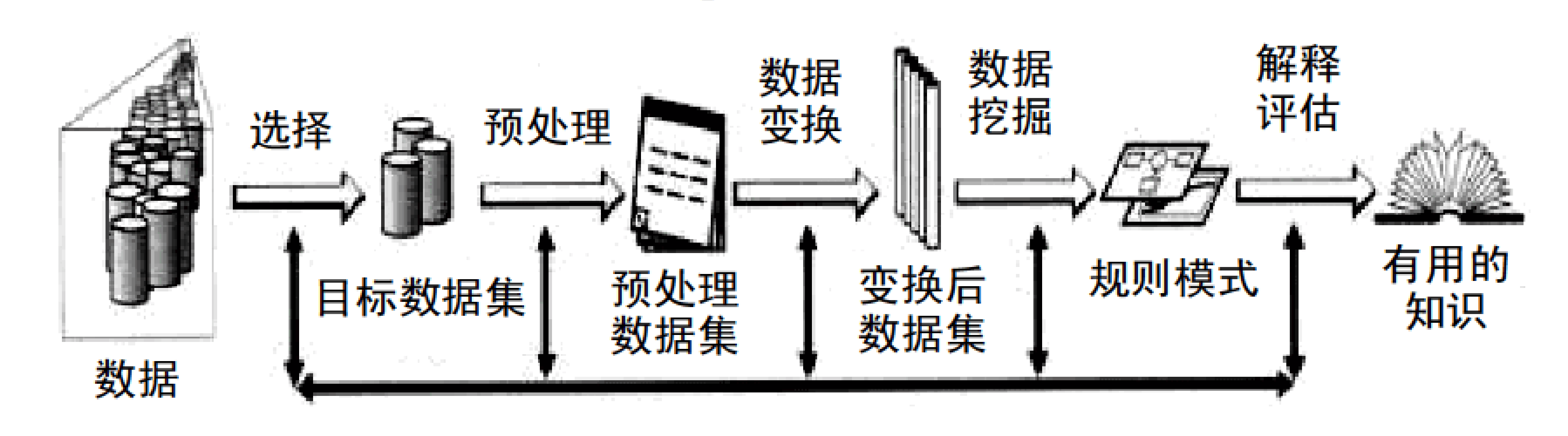


图 3-1　知识发现流程（The KDD Process）

数据来源：Fayyad et al.（1996）

其流程步骤是：先理解要应用的领域，熟悉相关知识，接着建立目标数据集，并专注所选择（selection）数据子集；再从目的数据中做前置处理（pre-processing），去除错误或不一致的数据；然后作数据简化与转换工作（transformation）；再经由数据挖掘的技术程序生

成为模式（patterns）、做回归分析或找出分类型态；最后经过解释/评估（interpretation/evaluation）成为有用的知识。这些程序是一个循环的关系，一直重复的步骤，最后才得到一些有用的知识。所以，KDD 是一连串的程序，数据挖掘只是其中的一个步骤而已。

3.4　OLAP 与数据挖掘的关系

所谓 OLAP（online analytical process），是指由数据库所链接出来的在线查询分析程序。简单地说，OLAP 是由使用者所主导，使用者先有一些假设，然后利用 OLAP 来查证假设是否成立；而数据挖掘则是用来帮助使用者产生假设。所以在使用 OLAP 或其他 query 的工具时，使用者是自己做探索（exploration），但数据挖掘是用工具帮助做探索。所以可以认为：数据挖掘用于产生假设，OLAP 则用于查证假设。

数据挖掘常能挖掘出超越归纳范围的关系，可以找出甚至不会被怀疑过的数据型样与关系的特性，事实上已超越了人们经验、教育、想象力的限制。而 OLAP 仅能利用人工查询及可视化的报表来确认某些关系。OLAP 与数据挖掘的比较如表 3-3 所示。

表 3-3　OLAP 与数据挖掘比较

在线分析处理（OLAP）	数据挖掘（data mining）
公司邮寄广告顾客回复率多少	哪些顾客容易回复公司的邮寄广告
新产品销售与客户数量	何种类型的老客户较倾向购买公司新产品
公司上年度十大客户	公司上年获利度最高的十大客户
哪些客户上个月并未续约	哪些客户较可能在未来的半年中不再续约
哪些客户贷款逾期未付	哪些客户贷款较易逾期支付
上一季度地区性销售报告	明年各地区产品的预测销售收入
昨日生产线次品率	如何提高产品的合格率

数据来源：Noonan 2000

3.5　数据挖掘与机器学习的关系

机器学习这门学科所关注的问题是：计算机程序如何随着经验积累自动提高性能？近年来，机器学习被成功地应用于很多领域，从检测信用卡交易欺诈的数据挖掘程序，到获取用户阅读兴趣的信息过滤系统，再到能在高速公路上自动行驶的汽车。同时，这个学科的基础理论和算法也有了重大的进展。

在数据挖掘领域，机器学习是相当重要的组成部分。机器学习中的大量算法都被用于大型数据库的探索分析和模式识别。例如：决策树学习算法已经被美国国家航空和航天局（NASA）用来分类天体，数据来自第二帕洛马天文台太空调查（Fayyad et al，1995）。这一系统现在被用于自动分类太空调查中的所有天体，其中包含了 3TB 的图像数据。

机器学习在很多应用领域被证明有很大的实用价值。它们在以下方面特别有用：

（1）数据挖掘问题，即从大量数据中发现可能包含在其中的有价值的规律，例如，从患者数据库中分析治疗的结果，或者从财务数据中得到信用贷款的普遍规则；（2）在某些困难的领域中，人们可能还不具有开发出高效的算法所需的知识，例如，从图像库中识别出人脸；（3）计算机程序须动态地适应变化的领域，例如原料供给在变化的环境下自动进行生产过程控制。

3.6 网络挖掘与数据挖掘的关系

网络挖掘（web mining）可以看做数据挖掘应用在网络数据的泛称。利用数据挖掘技术可以进行深入的网络访问数据分析，并建立精准的预测模型，实现智能化的个人网络服务。

网络挖掘除了统计对网页浏览率以及访客人次等日志文件的分析外，只要经由网络上的商品零售、财务服务、通信服务、医疗咨询、远距教学等由网络传送的数据，都可以归入到网络挖掘的范围，甚至可以整合 off-line 数据和 on-line 数据，实施更大规模的模型预测与估计。凭借因特网的便利性与渗透力，再借助网络行为的可追踪性与高互动性，网络挖掘完全有可能成为实现一对一营销理念的最佳技术和工具。

整体而言，网络挖掘具有以下特性：

- 数据收集容易且不引人注意。网络用户进入网站后的一切浏览行为与过程都可随时记录。
- 以交互式个性化服务为终极目标。除不同访客显示定制的网页外，对于不同的网络用户也应该提供不同类型的浏览服务。
- 可整合其他 off-line 数据，让网络挖掘的功能发挥更为充分。

第 4 章　数据挖掘商业软件产品及其应用现状

4.1　数据挖掘商业软件的分类

数据挖掘工具的软件市场大致可分为三类。

1．通用分析目的的软件包

SQL 2005
SAS Enterprise Miner
IBM Intelligent Miner
Unica PRW
SPSS Clementine
SGI MineSet
Oracle Darwin
Angoss KnowledgeSeeker

2．针对特定功能或行业而研发的软件包

KD1（针对零售业）
Options & Choices（针对保险业）
HNC（针对信用卡欺诈或呆账检测）
Unica Model 1（针对营销业）

3．整合 DSS/OLAP/Data Mining 的大型分析系统

Cognos Scenario and Business Objects

4.2　主要软件的介绍

以下介绍一般常用的数据挖掘工具的分类，如表 4-1 所示。

表 4-1　常用数据挖掘工具

分 析 工 具	定　　义	代表性产品
case-based reasoning	在关系型数据库中提供一个 means 找出 record 以发现类似规范的记录或一般记录	❑ CBR Express ❑ Esteen ❑ Kate-CBR ❑ The Easy Reasoner

续表

分 析 工 具	定 义	代表性产品
data visualization	其目标是从不同的角度，让信息以图形方式显示，使用户容易和快速地使用。此工具把不同数据层次加以集合或汇总，让用户快速地了解	❑ Alterian ❑ AVS/Express ❑ Visualization Edition ❑ Axum ❑ Discovery ❑ SPSS Diamond ❑ Visual Insight
fuzzy query and analysis	模糊理论积极承认人的主观性问题的存在，进而以模糊集合来处理不易量化的问题，故能找出意想不到的信息	❑ CubiCalc ❑ FuziCalc ❑ Fuzzy TECH for business ❑ Quest
knowledge discovery	这些工具特别设计以便确认那些已存在变量间的显著关系，也就是当它们可能有多重关系时，特别有用。这些数据挖掘工具能帮助指出庞大变量间的关系，发现盲点，创造巨大的商机	❑ Aria ❑ Answer tree ❑ CART ❑ DARWIN ❑ Enterprise Miner ❑ DataEngine
neural networks	类神经网络技术的目标是发现与预测数据的关系，与传统统计方法的区别是，它可以训练学习发现的关系，而且可适用于线性与非线性的情况，并可以弥补数据质量较差的情况，而处理出质量不错的信息来	❑ BackPack ❑ BrainMaker ❑ Loadstone ❑ NeuFrame/NeuroFuzzy ❑ Neural network Browser ❑ Neural connection ❑ Neural network Utility ❑ Neuralyst For Excel

4.3 顾客关系管理

顾客关系是指组织与其顾客间存在的各种互动关系。顾客关系管理（CRM）不仅可以提升企业与顾客间的互动关系，同时也可以通过互动关系来搜集顾客数据。

顾客关系管理并非信息科技，因此企业主应该了解在寻找合适的顾客关系管理软件的过程中，着重于已有顾客关系管理层面的考虑，而非寻找顾客关系管理的解决方案。因为任何一种顾客关系管理软件都不可能彻底解决企业与顾客间关系的维系与建立。完整的CRM 运作机制在相关的硬软件系统能够提供全面完善的支持之前，都有太多的数据准备工作与分析工作需要进行。企业通过数据挖掘可以分别对策略、目标定位、操作效能与测量评估四个方面的相关问题，有效率地从市场和客户搜集累积的数据中挖掘出对客户而言最关键、最重要的答案，从而建立真正的以客户需求为出发点的客户关系管理。

数据挖掘应用于 CRM 的主要方式对应于缺口分析（gap analysis）有三个部分：

① 针对客户获取的缺口（acquisition gap），可利用客户档案（customer profile）找出客户的一些共同特征，并深入了解客户，通过聚类分析对客户进行分群后再利用模式分析预测哪些人可能成为客户，帮助营销人员找到正确的营销对象，进而降低成本，也提高了营销的成功率。

② 针对销售提升的缺口（sales gap），可利用购物篮分析客户的消费特征，找出哪些产品是客户最容易一起购买的，或是利用序列分析预测客户在买了某产品后，在多久之内会买另一产品等。利用数据挖掘可以帮助企业定制更为有效的商品组合、产品推荐、进货量或库存量，甚至是如何摆设货架和商品等，同时也可以用来评估促销活动的成效。

③ 针对客户保留的缺口（retention gap），在商业竞争中，常会看到一些客户从原来的商家转入到其竞争对手的商家。通过分析这些转移的客户群资料，得出客户流失的基本特征，就可以在现有客户群中识别出可能转向的客户，然后设计一些保留措施以预防客户流失。

4.4　数据挖掘的行业应用

有关数据挖掘的行业应用如表 4-2 所示。

表 4-2　数据挖掘的行业应用

行 业 领 域	具 体 应 用
信用卡业	信用卡公司可使用数据挖掘来设定信用卡额度、购买授权决定、分析持卡人的购买行为、检测诈骗行为等
零售业	利用销售数据，实施促销活动，或评测广告宣传的效果；利用购物篮分析来了解顾客购买行为和偏好
金融业	证券分析师广泛使用数据挖掘来分析大量的财务数据以建立交易及风险模型来发展投资策略
银行业	搜集并分析详细的顾客信息，然后整合为营销策略；使用数据挖掘以识别顾客的贷款活动、定制金融产品；进行客户管理以寻找新的客户及加强客户忠诚度
直销业	使用数据挖掘可节省运营成本并且能够精确取得目标顾客、减少通话数量，且可以增加成功通话的比率；利用数据挖掘分析顾客群的消费行为与交易记录，结合基本数据，实现市场分割
制造业	数据挖掘已经广泛应用于制造业的流水线设计。例如，使用数据挖掘来检测潜在的质量问题，减少次品
电信业	使用数据挖掘，电信公司可以提供给顾客符合其需求的定制服务
保险业	利用数据挖掘技术来发现新的投保客户，减少客户流失，还可以有效检测保险欺诈



第 2 篇

Excel 2007 数据挖掘模块介绍

- 安装与设定 Excel 2007 数据挖掘加载项
- Excel 2007 数据挖掘入门
- 决策树
- 贝叶斯概率分类
- 关联规则
- 聚类分析
- 时序聚类
- 线性回归
- Logistic 回归
- 类神经网络
- 时间序列分析
- DMX 介绍

第 5 章　安装与设定 Excel 2007 数据挖掘加载项

5.1　系 统 需 求

在安装 Excel 2007 数据挖掘加载项前，需要了解相关系统的软硬件配置标准。其配置如下：

- 操作系统：Windows XP SP2、Windows Vista、Windows 2000 SP4、Windows 2003 SP1。
- Excel 2007：Professional、Professional Plus、Ultimate、Enterprise。
- 硬盘空间：至少 40MB 可使用空间。
- SQL Server 2005：SP1、SP2、RTM。注意：在同一台计算机上，安装数据挖掘加载项与 SQL Server 2005 时，SQL Server 2005 SP2 的 CTP（Community Technology Preview）版本与数据挖掘加载项是无法正常运作的。
- SQL Server 2005 Analysis Services：安装数据挖掘加载项必须连接 SQL Server 2005 Analysis Services 才能正常运行。支持 Analysis Services 的 SQL 2005 版本有：Enterprise Edition SP1、SP2、RTM，Standard Edition SP2。
- .NET：Microsoft .NET Framework 2.0。
- 旧版本删除：若在 2007 年 3 月 21 日前安装过 Office 2007 数据挖掘加载项，则必须删除后再重新安装。

5.2　开 始 安 装

数据挖掘加载项安装文件可从微软官方网站的下载中心下载，下载网址为 http://www.microsoft.com/downloads/details.aspx?displaylang=zh-cn&FamilyID=7c76e8df-8674-4c3b-a99b-55b17f3c4c51。

Step1：双击 SQLServer2005_DMAddin.msi 图标。

Step2：弹出【欢迎使用 SQL Server 数据挖掘外接程序安装向导】窗口，如图 5-1 所示，单击【下一步】按钮。

Step3：弹出许可协议窗口，如图 5-2 所示，选中【我同意许可协议中的条款】单选按钮，单击【下一步】按钮。

Step4：在如图 5-3 所示的【注册信息】窗口中输入姓名及公司名称，单击【下一步】按钮。

Step5：在如图 5-4 所示的【功能选择】窗口中，分别右击【Excel 数据挖掘客户端】

和【Visio 数据挖掘模板】选项，在弹出的快捷菜单中选择【安装此功能到本地硬盘上】命令，图标 就会变为 ，单击【下一步】按钮。



图 5-1 【欢迎使用 SQL Server 数据挖掘外接程序安装向导】窗口



图 5-2 许可协议窗口

Microsoft SQL Server 2005 数据挖掘外接程序安装程序

注册信息

在下面的字段中输入您的姓名以及单位名称。

姓名:

公司:

< 上一步(B)　下一步(N) >　取消

图 5-3 【注册信息】窗口



图 5-4　【功能选择】窗口

Step6：弹出如图 5-5 所示的【准备安装程序】窗口，单击【安装】按钮。

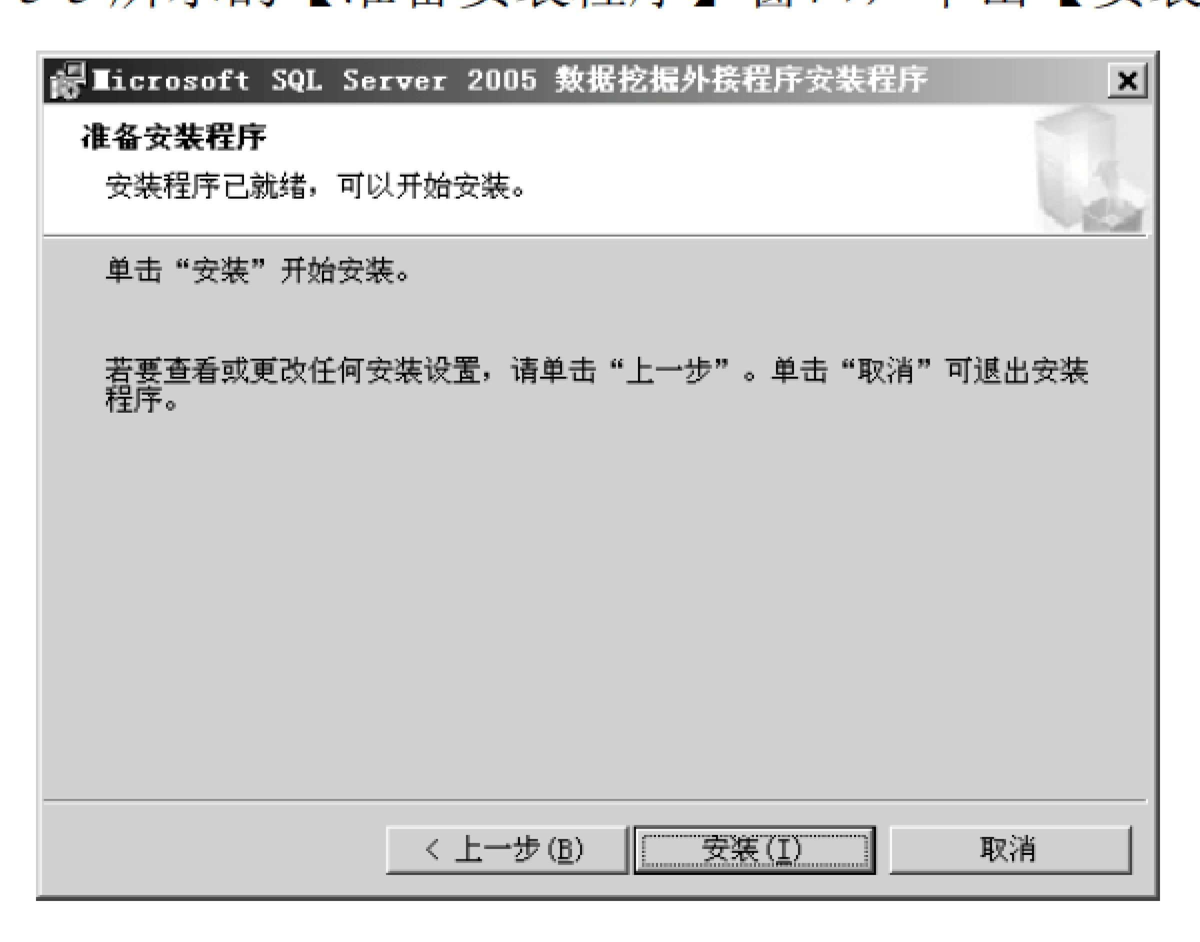


图 5-5　【准备安装程序】窗口

Step7：当【完成】按钮为可选状态时，说明已经安装完成了，如图 5-6 所示。



图 5-6　安装完成

5.3 完成安装验证

安装完成后的数据挖掘加载项，可从【开始】菜单的【所有程序】中找到新增加的【Microsoft SQL Server 2005 数据挖掘加载项】。安装的功能选择默认有：Excel 数据表分析工具、服务器组件公用程序。因为我们在安装时选择安装所有功能，所以会出现以下的功能：

- ❑ Data Mining Visio Template。
- ❑ 服务器组件公用程序。
- ❑ 开始。
- ❑ 说明和文件集。
- ❑ 范例 Excel 数据。

5.4 组 件 设 定

在使用数据挖掘加载项前，必须先确认是否已经连接设定到 SQL Server 2005 Analysis Services 数据库。连接设定的方式可以选择【服务器配置实用工具】或者【开始】命令，其操作过程类似，这里如图 5-7 所示选择【开始】命令。



图 5-7 选择【开始】命令

Step1：执行【开始】命令。

Step2：选择要连接的 SQL Server 2005 Analysis Services 实例，这里选中第二个单选按钮，因为要连到本机的 Analysis Services 数据库。单击【下一步】按钮，如图 5-8 所示。

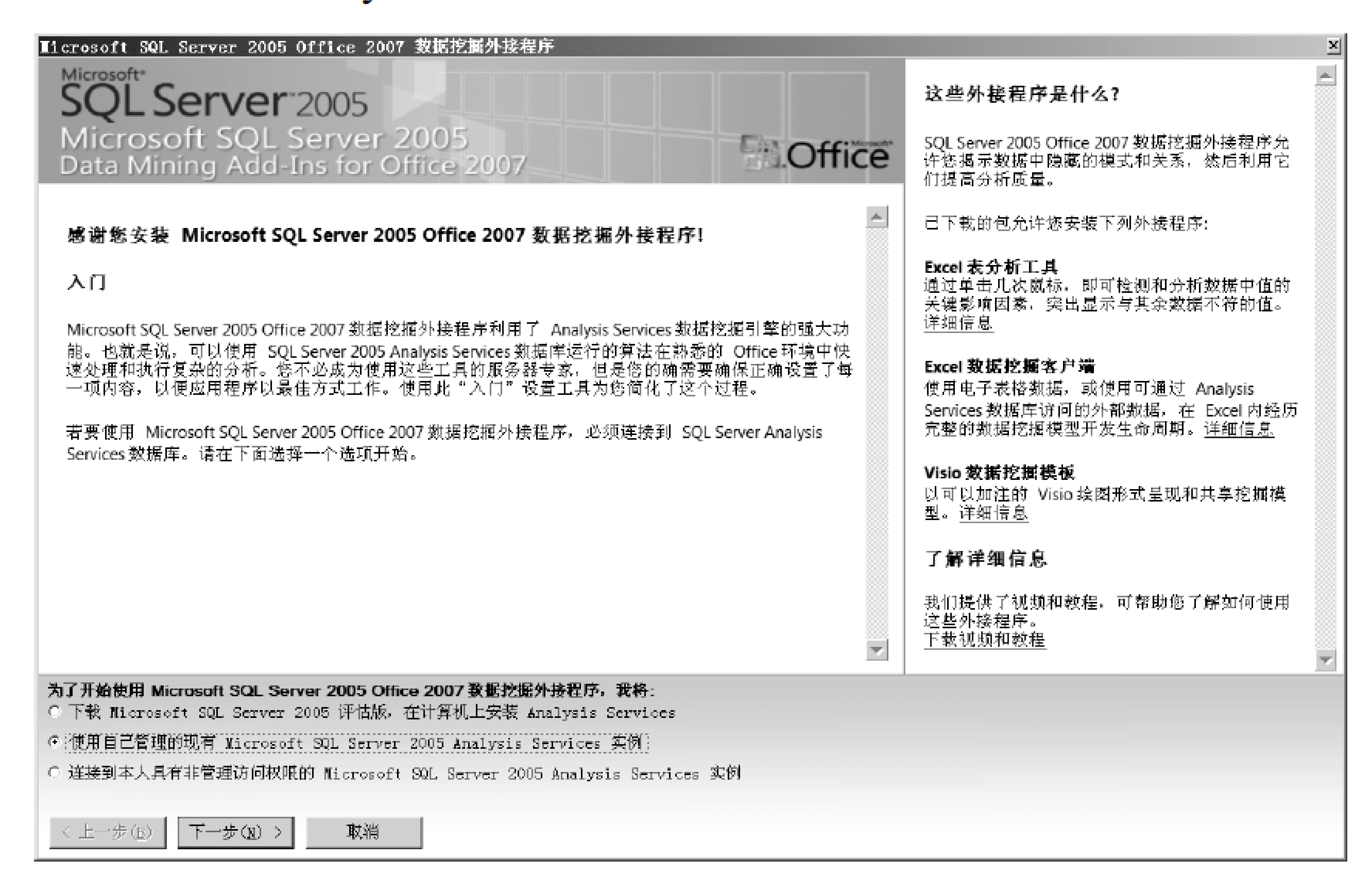


图 5-8 选择要连接的 SQL Server 2005 Analysis Services 实例

Step3：执行服务器配置实用工具，数据挖掘加载项安装后的配置工具目录 Microsoft. SqlServer.DataMining.Office.ServerConfiguration.exe 会保存在 C:\Program Files\Microsoft SQL Server 2005 DM Add-Ins 数据夹内。选择该程序以运行连接设定，如图 5-9 所示。



图 5-9　运行服务器配置工具

Step4：这里相当于重新选择【服务器配置实用工具】命令。开始进入数据挖掘加载项配置向导设定，单击【下一步】按钮，如图 5-10 所示。



图 5-10　【欢迎使用 SQL Server 2005 数据挖掘外接程序配置向导】窗口

Step5：输入要连接的 Analysis Services 数据库服务器名称，因为要连到本机，故在【服

务器名称】文本框中输入“localhost”，如图 5-11 所示。单击【下一步】按钮。



图 5-11　输入服务器名称

Step6：弹出【正在连接到服务器 localhost...】窗口，如图 5-12 所示。



图 5-12　【正在连接到服务器 localhost...】窗口

若 SQL Server Analysis Services 未运行，则会出现无法连接服务器的信息 无法连接到服务器“localhosts”。请确保用户“SQLAI\Shengqiang_LAI”至少具有对服务器上某数据库的读取权限。，将 Analysis Services 服务启动后，单击【下一步】按钮。

Step7：是否要建立临时挖掘模型。所建立的临时挖掘模型会在断开连接后自动移除。当启用临时挖掘模型功能时，相应地会增加使用内存占用量及硬盘空间。若允许建立，则选中【允许创建临时挖掘模型】复选框；若不允许建立，则取消选中该复选框。单击【下一步】按钮，如图 5-13 所示。

Step8：建立数据库，这里是指数据挖掘加载项所使用的数据库，可以直接使用现有的数据库，或建立新的数据库。例如建立一个新的数据库名称为 DMAddinsDB-Test。在【数

据库名称】文本框中输入“DMAddinsDB-Test”，单击【下一步】按钮，如图 5-14 所示。

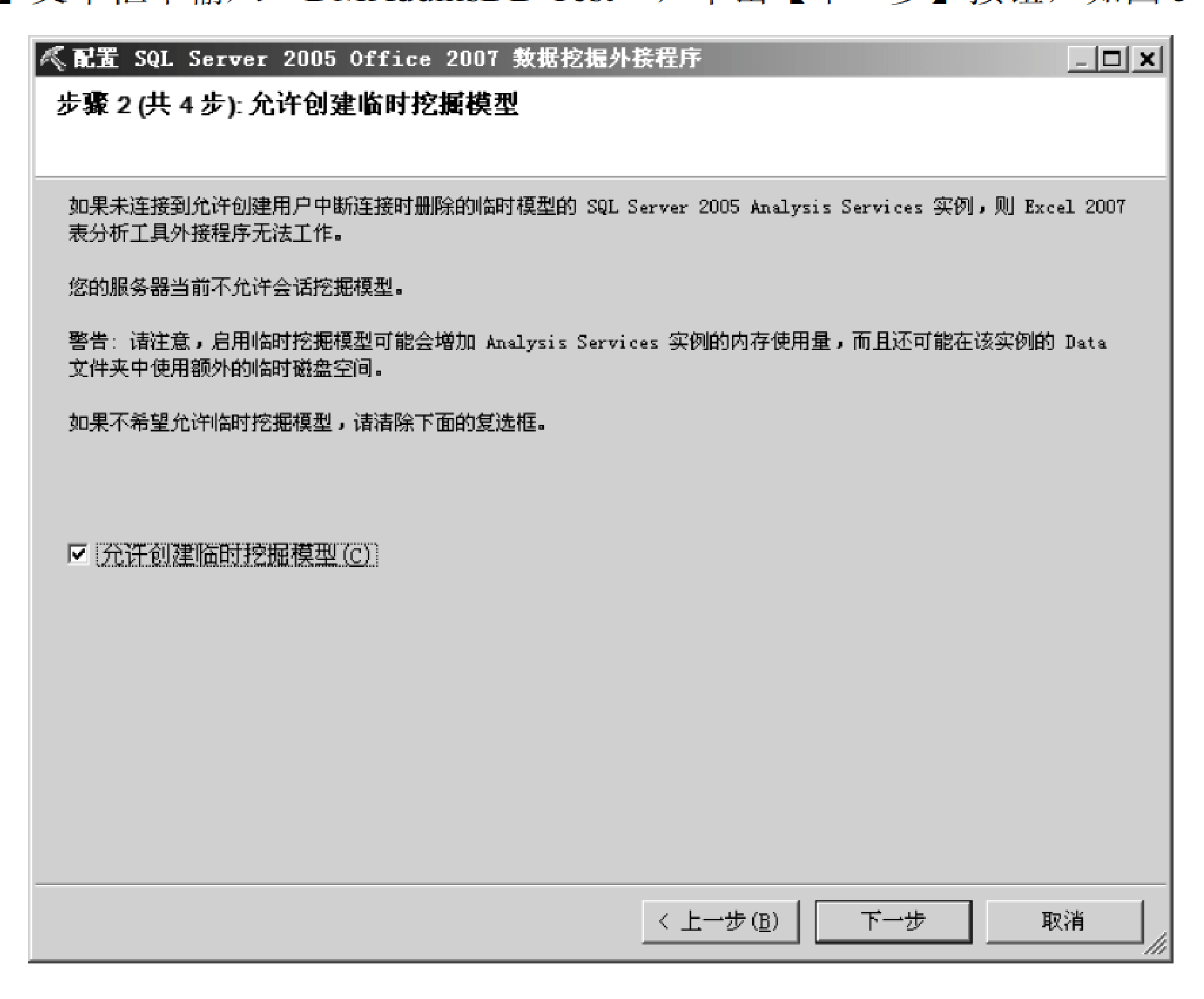


图 5-13　选中【允许创建临时挖掘模型】复选框



图 5-14　创建新数据库

Step9：用户的权限，授权使用数据挖掘加载项数据库。选中【将数据库管理权限授予

外接程序用户】复选框，此权限让使用者能够新增、修改、删除对象等。单击【完成】按钮，如图 5-15 所示。



图 5-15　授予用户权限

Step10：组件设定确认，如图 5-16 所示每一个设定动作确认成功，单击【关闭】按钮。



图 5-16　组件设定确认成功

5.5　配置完成检查

在 5.4 节成功地设定连接服务器以及新增的数据库，可在以下几个地方来查看。

1. SQL Server Management Studio

Step1：运行 SQL Server Management Studio，选择服务器类型为 Analysis Services，服务器名称为 localhost，单击【连接】按钮，如图 5-17 所示。



图 5-17　运行 SQL Server Management Studio

Step2：展开【数据库】，就会看到 5.4 节所建立的数据库名称 DMAddinsDB-Test，如图 5-18 所示。



图 5-18　【数据库】中新增 DMAddinsDB-Test

2. Excel 2007

启动 Excel 2007 后，在功能选单上会出现【数据挖掘】选项卡的功能，如图 5-19 所示。

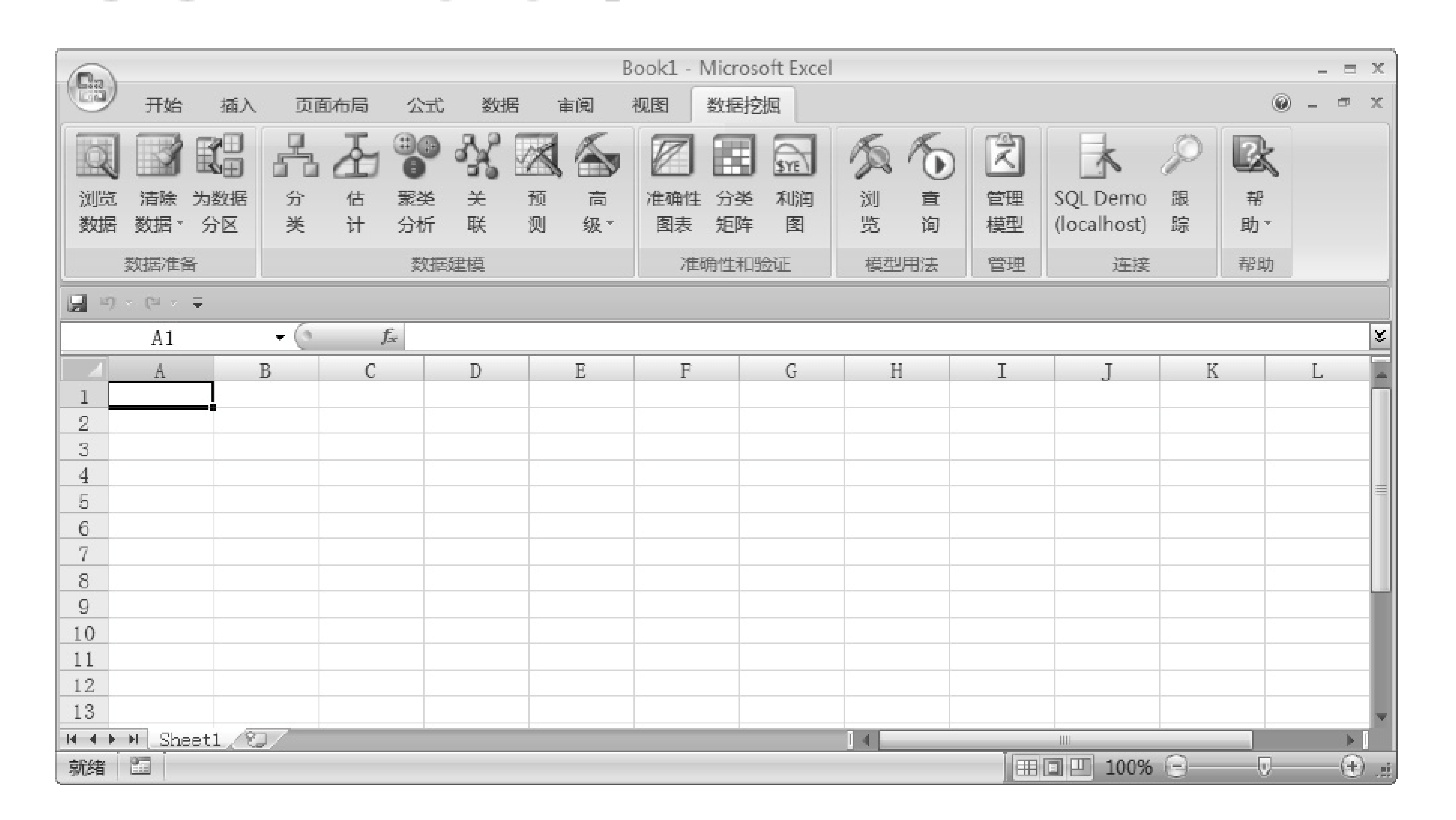


图 5-19　功能选单上出现【数据挖掘】选项卡

第 6 章　Excel 2007 数据挖掘入门

6.1　Excel 2007 数据挖掘功能介绍

Excel 2007 数据挖掘功能选项分成七大区块的工具栏如图 6-1 所示，七大区块功能介绍如下。



图 6-1　数据挖掘工具栏

- 数据准备：在开始数据挖掘之前，可先对数据做单一查看、清除整理数据或随机抽样数据。数据准备的方式有浏览数据、清除数据，以及为数据分区。
- 数据建模：开始进行数据挖掘步骤，可以建立挖掘模型、预测分析等。数据模型化的方法有分类、估计、聚类、关联、预测以及高级等。
- 准确性和验证：通过图型来查看挖掘模型。图型有准确性图表、分类矩阵和利润图。
- 模型用法：可对已构建好的挖掘模型条件式查询其结果。
- 管理：可对已构建好的挖掘模型管理其挖掘结构。
- 连接：设定与追踪 Analysis Services 的连接。
- 帮助：取得数据挖掘加载项的使用说明。

6.2　数据挖掘使用说明

Excel 2007 数据挖掘功能选项中的说明，是针对数据挖掘加载项的使用说明，而 Excel 2007 软件工具的说明是 Excel 2007 窗口最右边的小图标，两者是独立的。数据挖掘的使用说明功能除了提供在线查询的方式外，还有帮助向导，以及教学影片，非常方便使用者学习。

6.2.1　目录查询

不论是依目录查询还是从索引关键词查询，都是用户最熟悉的功能，如图 6-2 所示。



图 6-2　目录查询

6.2.2　开始功能

Excel 2007 数据挖掘说明中的开始功能与第 5 章安装与设定操作方式一样，如图 6-3 所示，请参考第 5 章的安装说明。



图 6-3　开始功能界面

6.2.3　视频和教学

单击“下载影片和教学课程”链接后会自动链接到微软公司的网页，内有教学影片，读者可以自行选择观看，如图 6-4 所示。



图 6-4　视频和教学网页

6.3　数据挖掘连接配置

设定连接数据挖掘服务器，必须设定连接到 Analysis Services 数据库。

6.3.1　设定目前的连接

其操作步骤如下：

Step1：单击此功能会开启 Analysis Services 连接设定，在第 5 章安装与设定中已经设

定好一个连接，因为已经预先建立好了一个 SQL Demo 的分析数据库，系统会自动将它设定为默认值，如图 6-5 所示。若要增加 Analysis Services 连接，单击【新建】按钮。

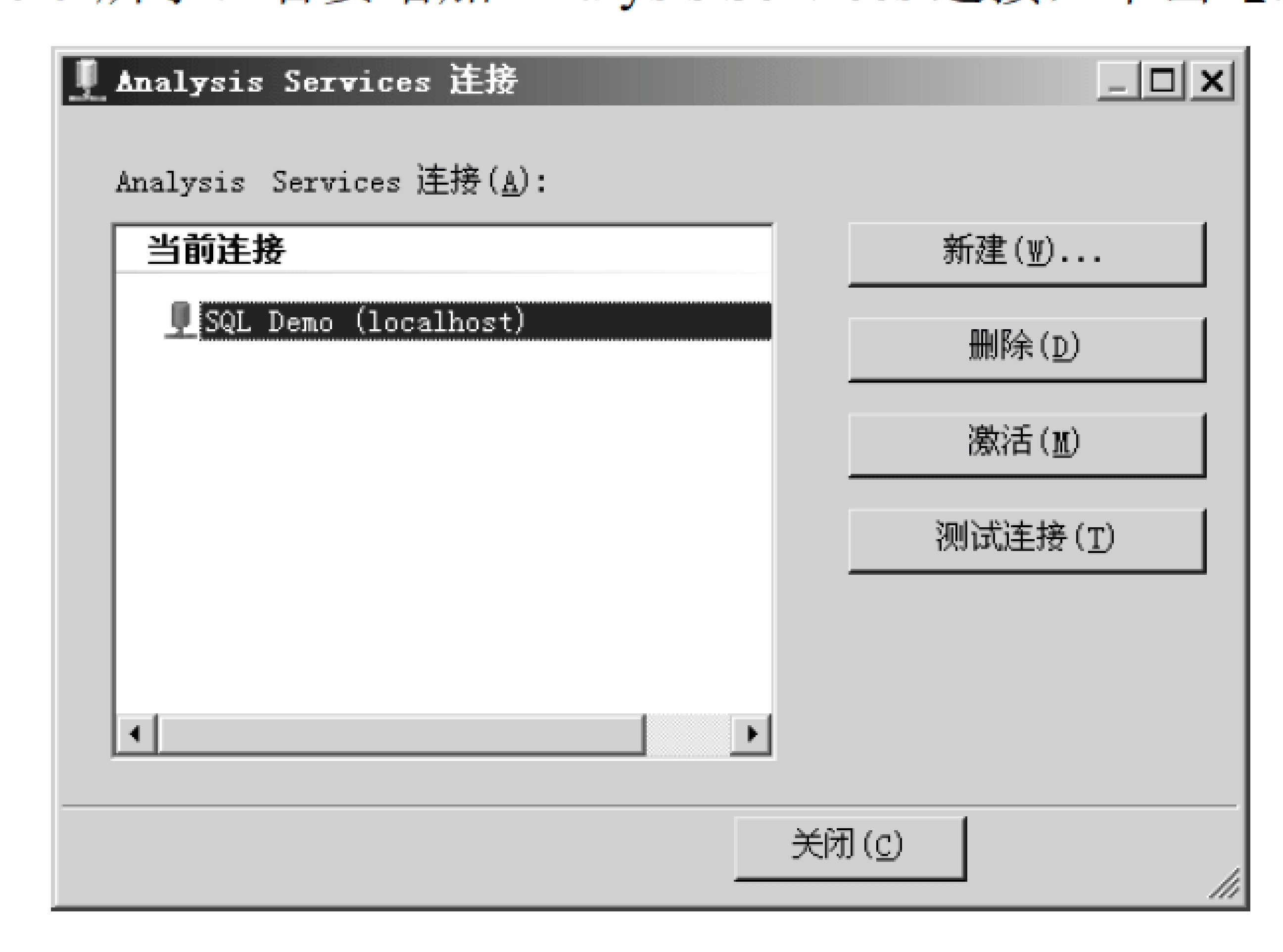


图 6-5 默认的连接

Step2：在【服务器名称】文本框中输入要连接的服务器名称：例如，localhost；

选择目录名称：例如，DMAddinsDB-Test；

输入容易记的名称：这里系统会将目录名称填入，也可以自行更改，例如，DMAddinsDB-Test（localhost）整个操作如图 6-6 所示。

Step3：在 Step2 中，可以单击【测试连接】按钮。出现【测试连接成功】提示框（如图 6-7 所示），表示已经新增连接完成，再单击两次【确定】按钮即可。

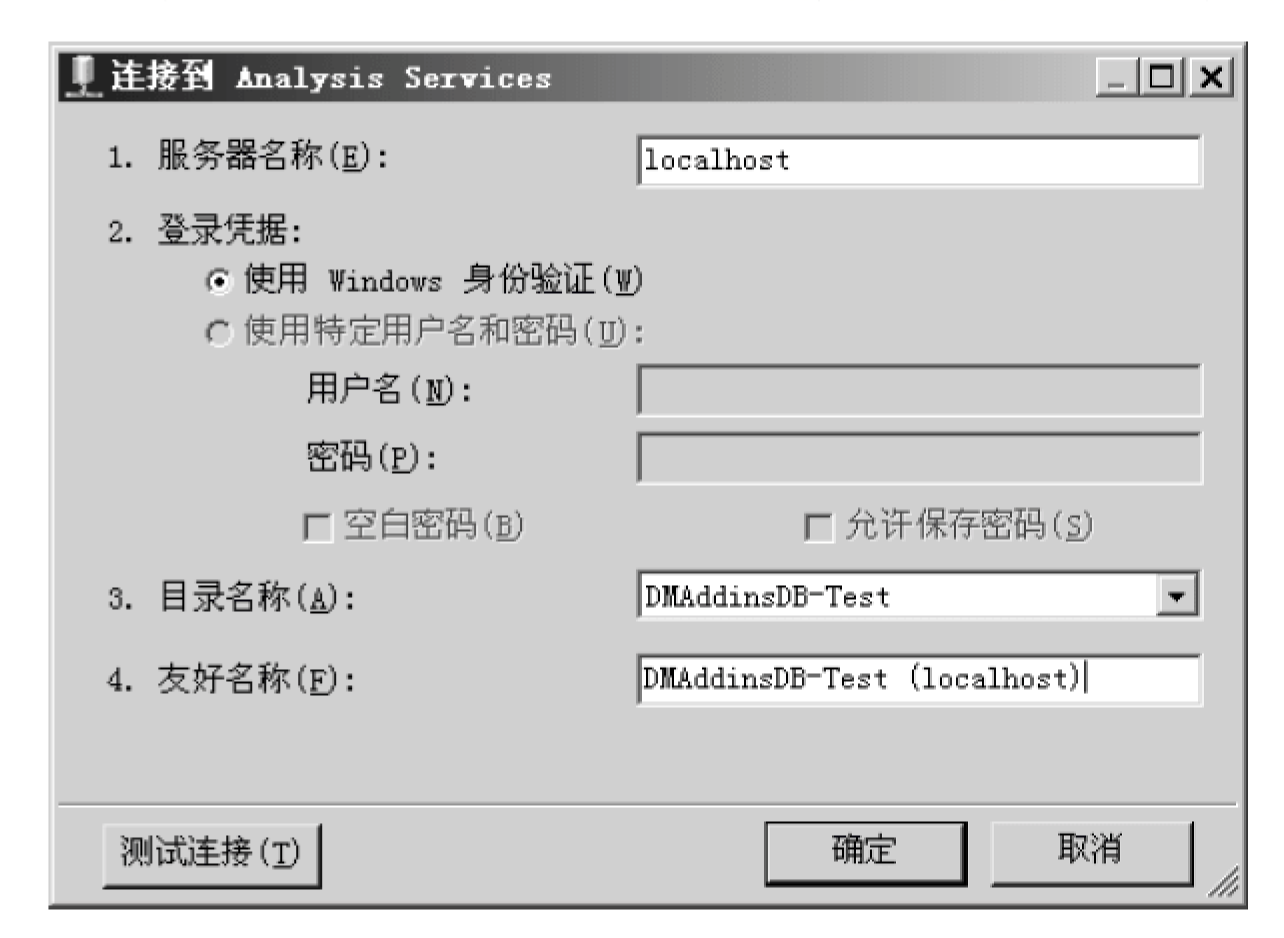


图 6-6 设定连接



图 6-7 测试连接成功

Step4：在目前的连接中增加了刚设定的名称，若要再变更连接，可在要连接的项目上双击，就会改变成目前的连接，如图 6-8 所示。然后单击【关闭】按钮。

Step5：在 Excel 2007 数据挖掘的连接功能上，可以看到已经改变连接了，如图 6-9 所示。



图 6-8　新增连接

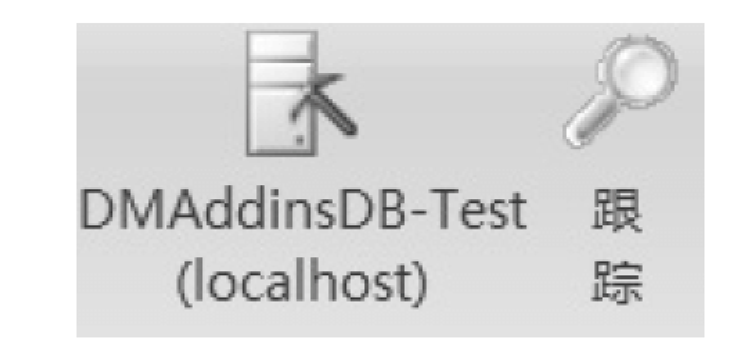


图 6-9　改变后连接

6.3.2　跟踪

此功能为跟踪传送到数据挖掘服务器的查询，选择当前连接就会显示连接查询，如图 6-10 所示。



图 6-10　跟踪器

6.4　数据准备

在开始数据挖掘之前，可先对数据做单一查看、清除整理数据或进行抽样。

6.4.1　浏览数据

浏览数据功能可以建立基本数据的统计信息，依据所选择的数据列产生直方图。操作步骤如下：

Step1：开始使用浏览数据向导，单击【下一步】按钮，如图 6-11 所示。

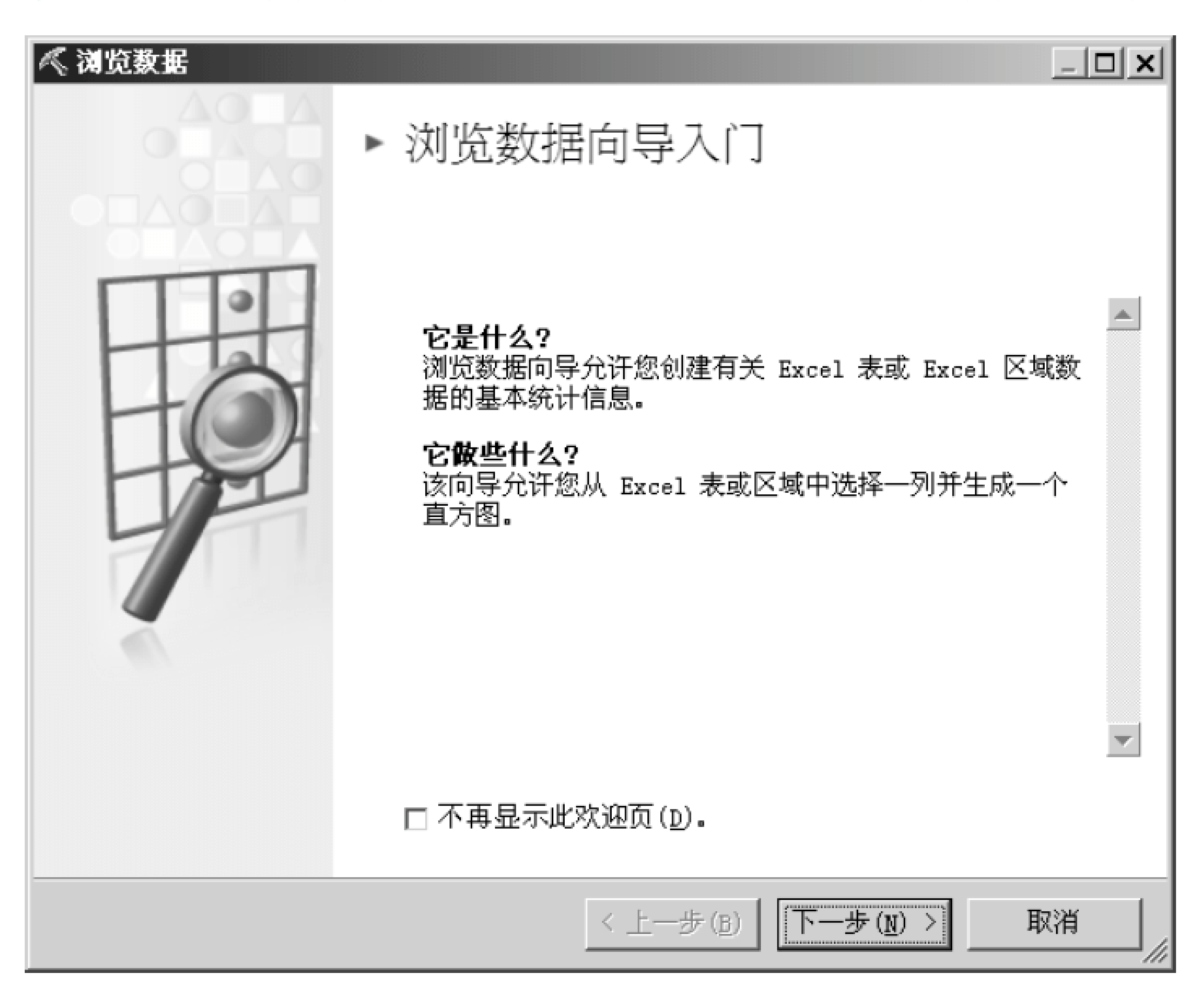


图 6-11 浏览数据向导

Step2：选择来源数据，选择数据表或设定数据区域，单击【下一步】按钮，如图 6-12 所示。



图 6-12 选择源数据

Step3：选择要分析的数据列，单击【下一步】按钮，如图 6-13 所示。

Step4：查看图。有两种方式查看。

① 以离散方式查看：无论数据为离散型或者是连续型，都可以用此图型查看，但是若

分析的数据为离散型数据（定性数据）时，只能以此图型查看，如图 6-14 所示。

浏览数据

选择列

选择列(S): Age

数据样本(D):

cupation	Home Owner	Cars	Commute Distance	Region	Age	Purchased Bike
ed Manual	Yes	0	0-1 Miles	Europe	42	No
cal	Yes	1	0-1 Miles	Europe	43	No
essional	No	2	2-5 Miles	Europe	60	No
essional	Yes	1	5-10 Miles	Pacific	41	Yes
cal	No	0	0-1 Miles	Europe	36	Yes
ual	Yes	0	1-2 Miles	Europe	50	No
agement	Yes	4	0-1 Miles	Pacific	33	Yes
ed Manual	Yes	0	0-1 Miles	Europe	43	Yes
cal	Yes	2	5-10 Miles	Pacific	58	No

< 上一步(B)　下一步(N) >　取消

图 6-13　选择列

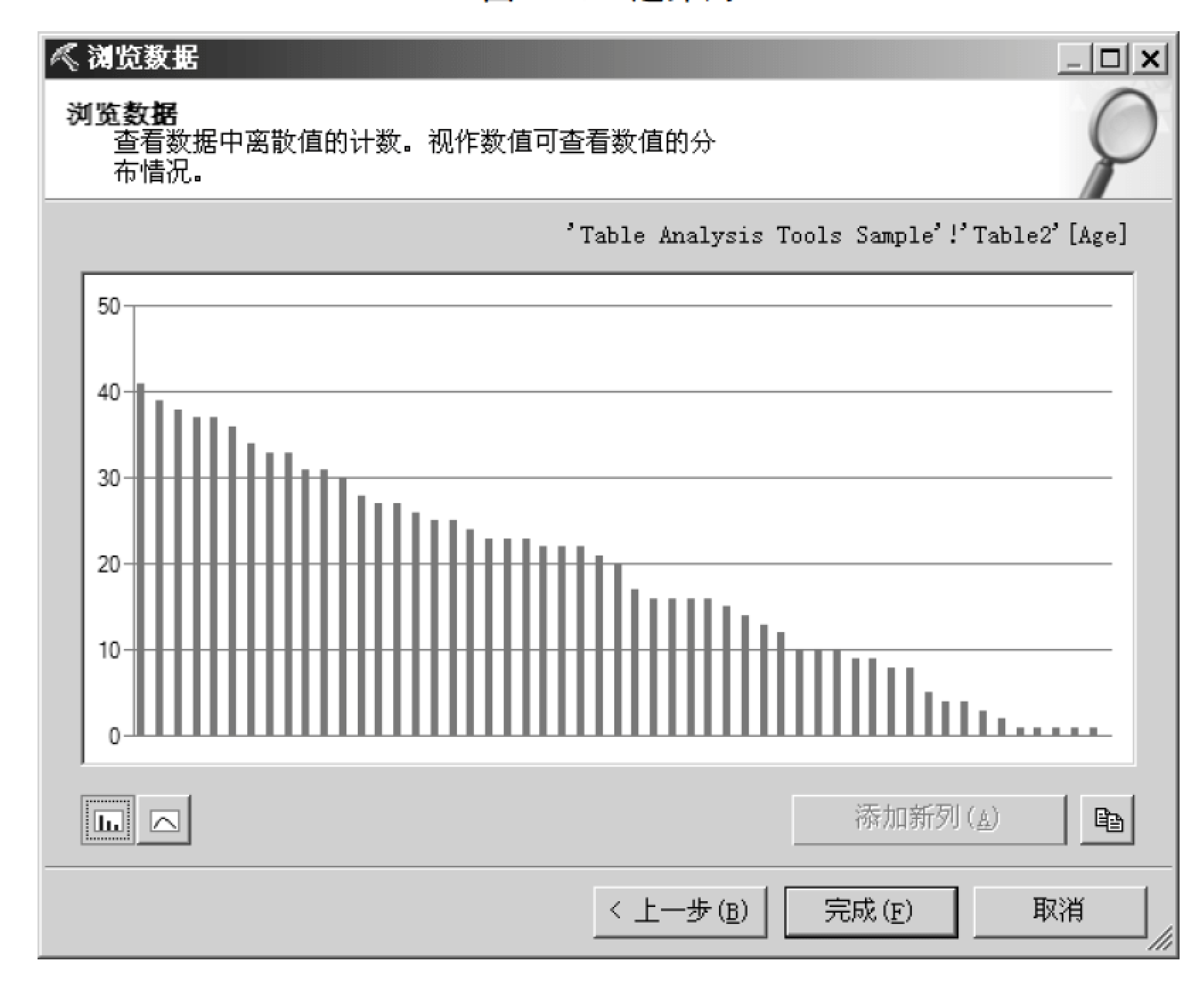


图 6-14　以离散方式查看

② 以数值方式查看：分析的数据为连续型数据时，可以用此图型查看。

- ❑ 存储桶：数据分组数，依据存储桶数字而定，如图 6-15 所示。
- ❑ 加入新数据列：依存储桶的分组数据，加到分析数据列的后面，单击【完成】按钮。在来源数据上，可以发现已经增加一列离散化后的数据，如图 6-16 所示。

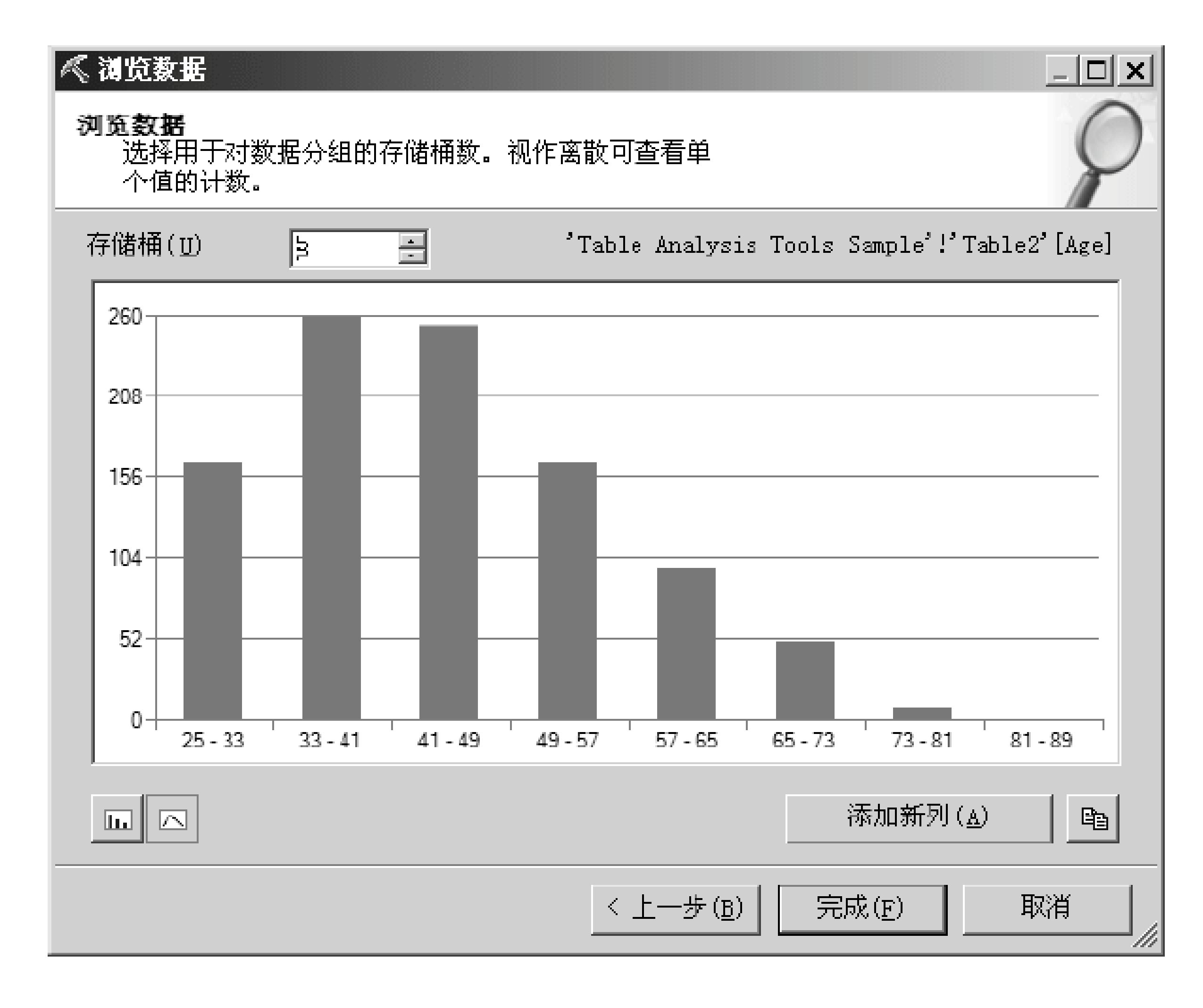


图 6-15　存储桶

Age	Age2
42	41 - 49
43	41 - 49
60	57 - 65
41	41 - 49
36	33 - 41
50	49 - 57
33	33 - 41
43	41 - 49
58	57 - 65
48	41 - 49
54	49 - 57
36	33 - 41
55	49 - 57
35	33 - 41
45	41 - 49
38	33 - 41
59	57 - 65
47	41 - 49

图 6-16　加入新数据列

6.4.2　清除数据

清除数据有两种，清除离群值数据与重新定义数据卷标。

1. 离群值

在分析数据的过程中，常会有一些数据超出正常范围，或者大大超出预期的范围，或是不正确的输入值等，这样的值都称为离群值。其操作 Step1～Step3 同浏览数据功能的操作步骤，不再赘述。

- 指定临界值：指定允许的范围，在范围值外其值会被移除。此例说明：年龄最大为 89 岁，最小为 25 岁，将年龄最大值设定到 65 岁时，则大于 65 岁以上的年龄区块会有阴影，阴影部分的观测将会被移除，如图 6-17 所示。

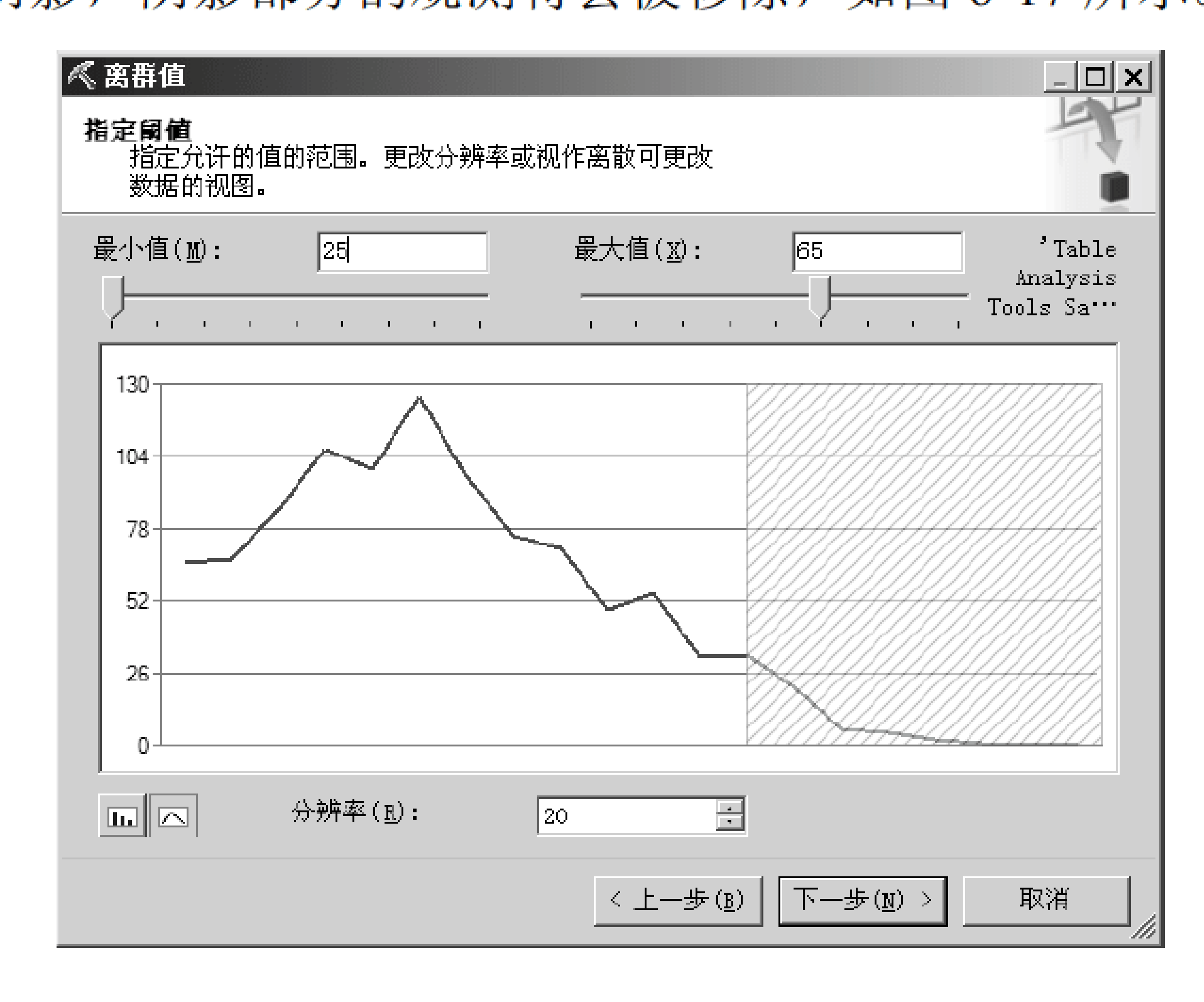


图 6-17　指定阈值

- 移除离群值的方式：指定一个移除离群值的方式，如图 6-18 所示。

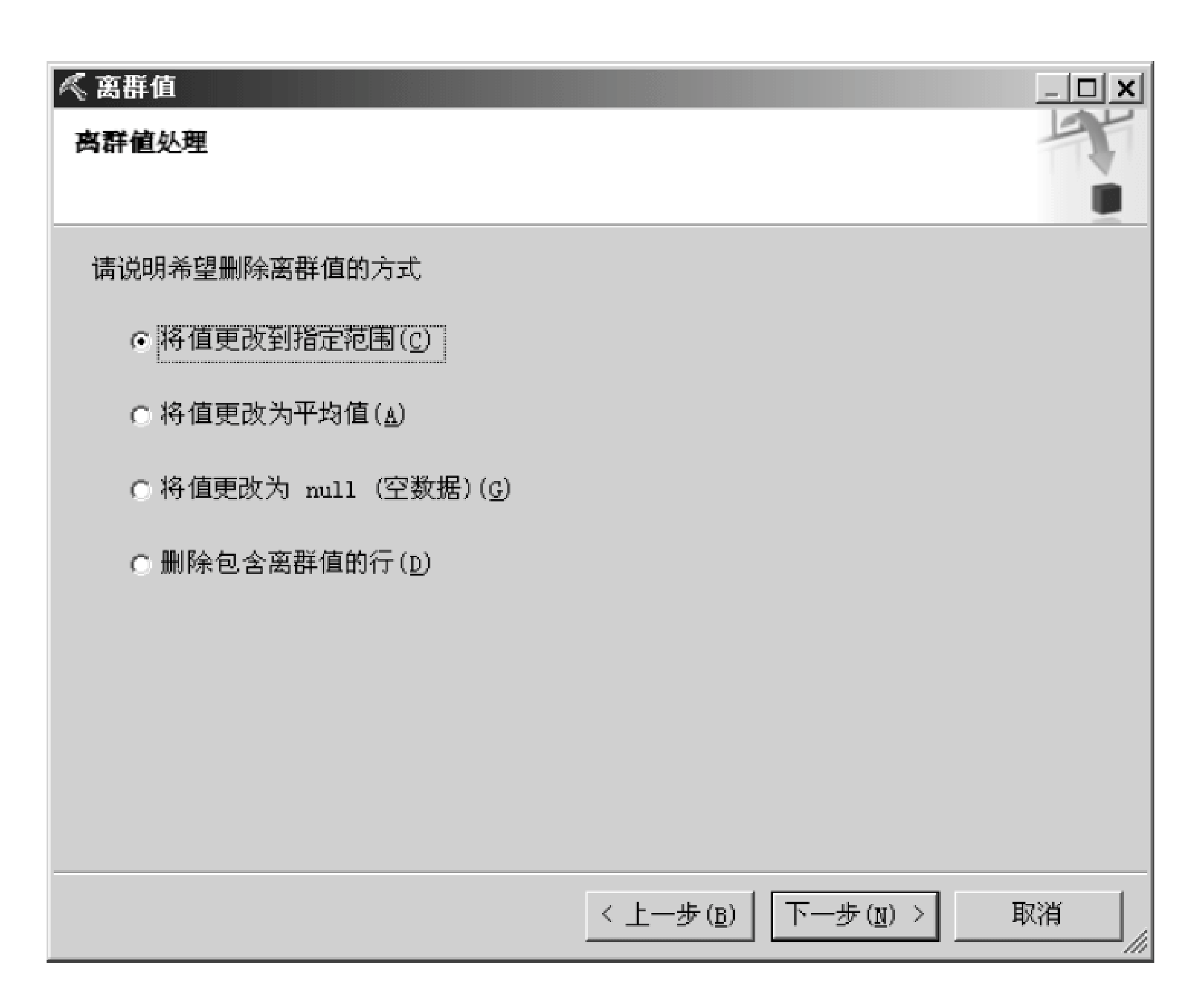


图 6-18　离群值处理

- ❑ 放置数据修改的位置：指定数据放置目的地，如图 6-19 所示。

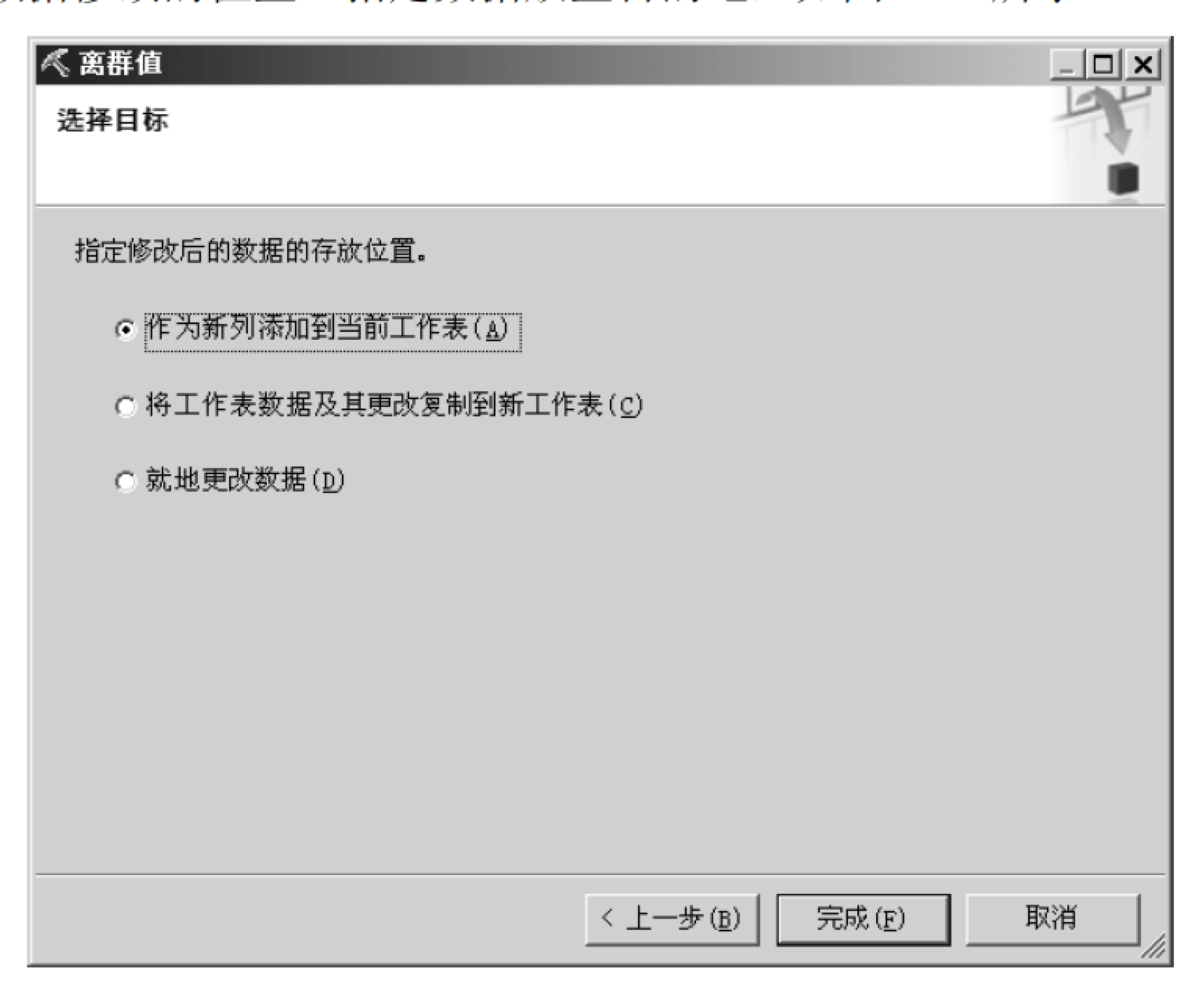


图 6-19　指定数据放置目的地

2. 重设标记

在分析数据的过程中，常会有一些数据的输入方式造成很难解释与解读，例如性别以数字 0、1 代表，此时就需要将数据列重新给定一个标签说明。其操作 Step1～Step3 同浏览数据功能的操作步骤，不再赘述。

- ❑ 给定新的标签：输入新的标签，如图 6-20 所示。



图 6-20　输入新的标签

- 放置数据修改的位置：指定数据放置目的地，如图 6-21 所示。



图 6-21　指定数据放置目的地

6.4.3　分割数据

数据挖掘前的数据抽样，有一个很重要的工作，就是要将数据分割为定型集（即训练数据集，training data set）与测试数据集（testing data set）。通常将来源数据的 70%作为训练数据集，来源数据的 30%作为测试数据集，比例并非固定，是可以调整的。其操作 Step1、Step2 同浏览数据功能的操作步骤，不再赘述。

数据抽样类型：选定数据抽样方式，如图 6-22 所示。

为数据分区向导

选择抽样类型

抽样方法

- 将数据拆分为定型集和测试集(S)
- 随机抽样(R)
- 过度抽样以调节数据分布(O)

说明

拆分数据根据您提供的比率将所选数据随机拆分为两个集，即定型集和测试集。定型集可以使用数据建模工具创建挖掘模型。模型创建后，测试集可以使用准确性和验证工具测试模型。每个集都在新工作表中创建。

< 上一步(B)　下一步(N) >　取消

图 6-22　选择抽样类型

1．将数据分割成定型集和测试集

依所提供的比例，将数据分割成定型集与测试集。定型数据集（训练数据集）用来构建数据挖掘模型，模型构建完成后，再将测试数据集通过准确性和验证工具进行测试验证。

（1）设定训练数据集的百分比，如图 6-23 所示。

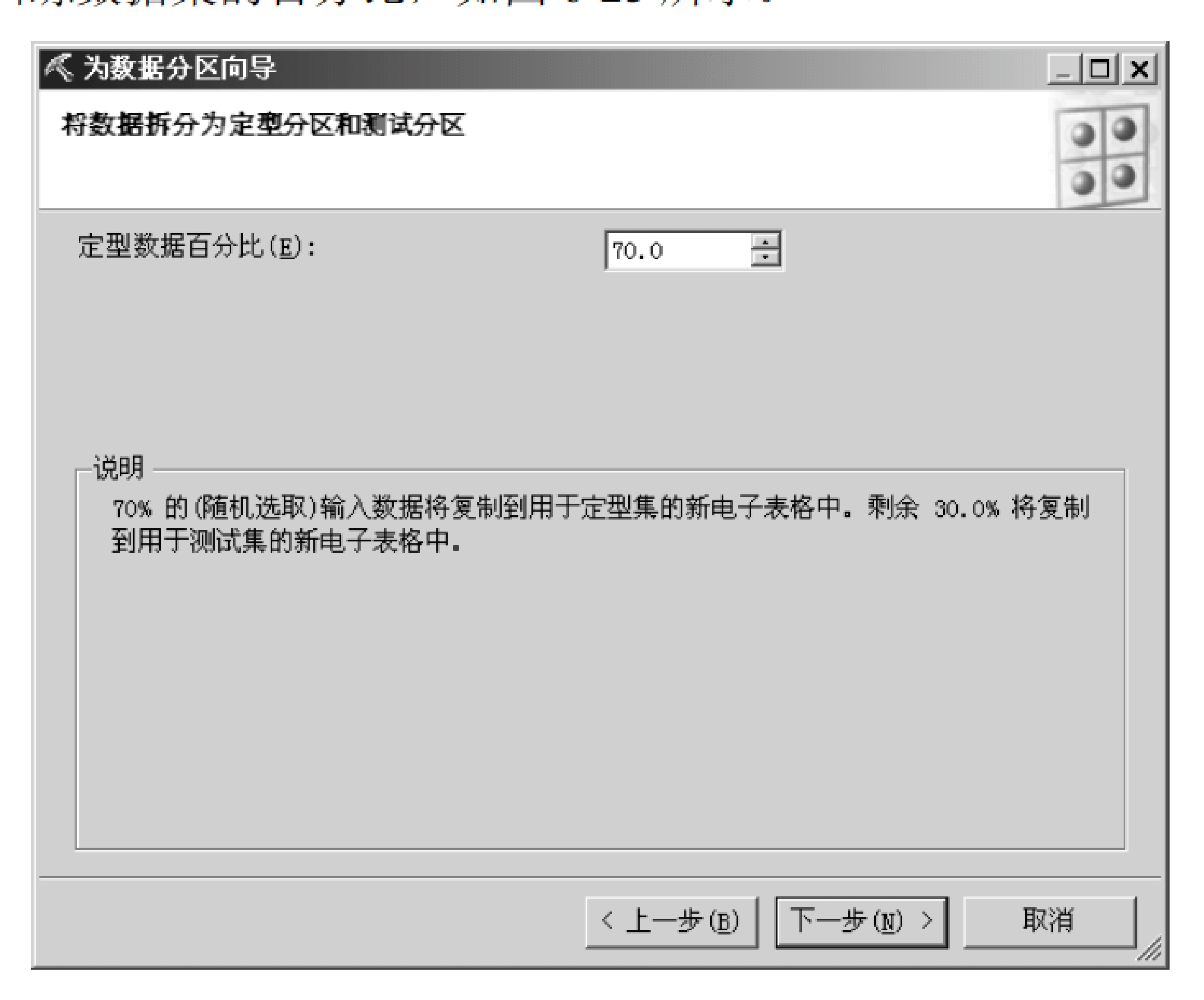


图 6-23　设定百分比

（2）输入训练数据集与测试数据集的工作表名称，如图 6-24 所示。



图 6-24　输入工作表名称

（3）增加了两个工作表：训练数据集与测试数据集，如图 6-25 所示。

Table Analysis Tools Sample　训练数据集　测试数据集

图 6-25　新增两个工作表

2. 随机抽样

以设定百分比或数目方式来抽样，而每个被选取的概率是相等的。被选取的数据会放置在新的工作表中，未选取的数据也可选择放置在另一个工作表中。随机抽样（Random Sampling）的方式可减少数据挖掘的数据量。另一种抽样方式是固定样本量，即设定样本的行计数。

（1）设定抽样的大小，如图 6-26 所示。



图 6-26　设定样本大小

（2）输入所选集工作表名称与未选集工作表名称，如图 6-27 所示。

为数据分区向导
完成
所选集工作表名称(E):　选取的数据集
☑ 为未选数据创建工作表(C)
未选集工作表名称(N):　未选取的数据集
< 上一步(B)　完成(F)　取消

图 6-27　输入工作表名称

（3）增加了两个工作表：选取的数据集与未选取的数据集，如图 6-28 所示。

Table Analysis Tools Sample　选取的数据集　未选取的数据集

图 6-28　在 Excel 2007 的下方标签行中新增两个工作表

3．超额抽样以平衡数据分布

超额抽样（oversampling）所建立的数据集中会包含以特定的超额比例选取的异常事件观测数据，若数据中正常观测数据和异常观测数据的比例差距较大时，可使两者的比例设定相当。不过由于样本偏误的因素，这一抽样方法较少使用。

（1）设定目标百分比，如图 6-29 所示。



图 6-29　设定目标百分比

（2）输入抽样数据工作表名称，如图 6-30 所示。



图 6-30　输入抽样数据工作表名称

（3）增加了一个工作表：抽样数据集，如图 6-31 所示。

Table Analysis Tools Sample　抽样数据集

图 6-31　在 Excel 2007 的下方标签行中新增一个工作表

6.5　数 据 建 模

下面开始构建数据挖掘模型。数据建模都是运用数据挖掘算法来构建模型，若是读者使用过 SQL Server 2005 的 Analysis Services 则会很熟悉。微软应用了数据挖掘的九个算法，在数据模型化中，除了列出常用的五个模型方法外，“高级”模型方法则是自行选择挖掘算法并以手动方式自行设定参数。

1．分类

依据分析的个体属性分类，算法为 Microsoft 决策树。

2．估计

依据模型相关的变量去预测一个连续型数据，算法为 Microsoft 决策树。

3. 聚类

将同质数据归为相同的类别，算法为 Microsoft 聚类分析。

4. 关联

发现所有相关程度较高的项目集合。算法为 Microsoft 关联规则。

5. 预测

根据分析个体属性的历史观察值预测未来值。算法有 Microsoft 时间序列和 Microsoft 决策树。

6. 高级

自行选择挖掘算法并以手动方式自行设定参数。

微软所提供的九种算法如下：

- Microsoft 决策树。
- Microsoft 贝叶斯概率分类。
- Microsoft 时序聚类。
- Microsoft 时间序列。
- Microsoft 聚类。
- Microsoft 线性回归。
- Microsoft Logistic 回归。
- Microsoft 关联规则。
- Microsoft 类神经网络。

以上九种数据挖掘算法与应用，将于后面章节分别介绍。

6.6 准确性和验证

数据挖掘模型构建完成后，可以通过准确性和验证方式，用图表查看模型的准确性。

6.6.1 准确性图表

使用查询中的测试数据，用来评估模型的效率。若模型的因变量是定性变量，则准确性以利润图显示；若因变量为数值型变量，则以散点图表显示。此功能可以建立基本数据的统计信息，依据所选择的数据列产生直方图，并将现有模型与假设理想模型做比较。分类模型和估计模型分别如图 6-32、图 6-33 所示。

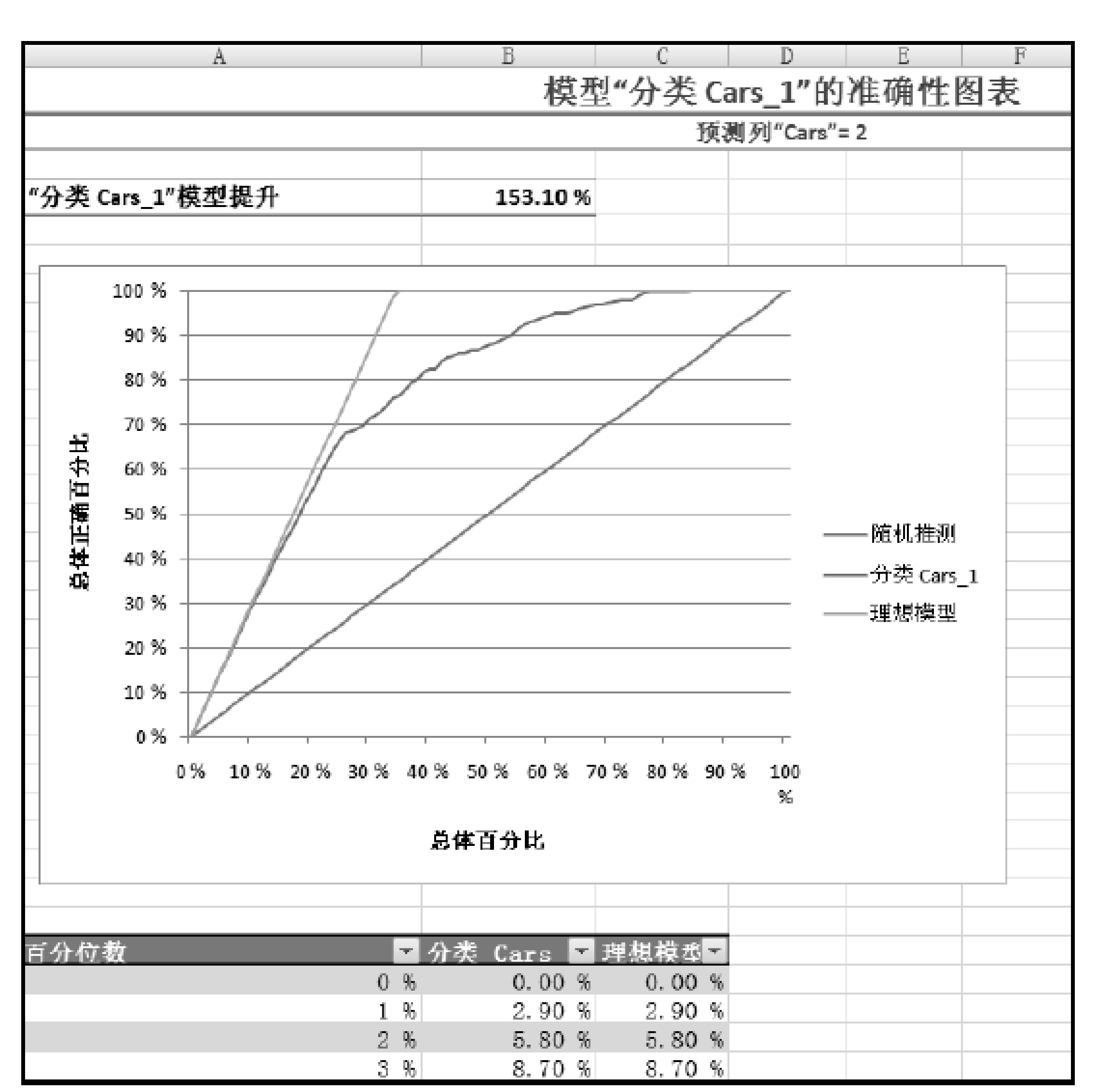


图 6-32　分类模型

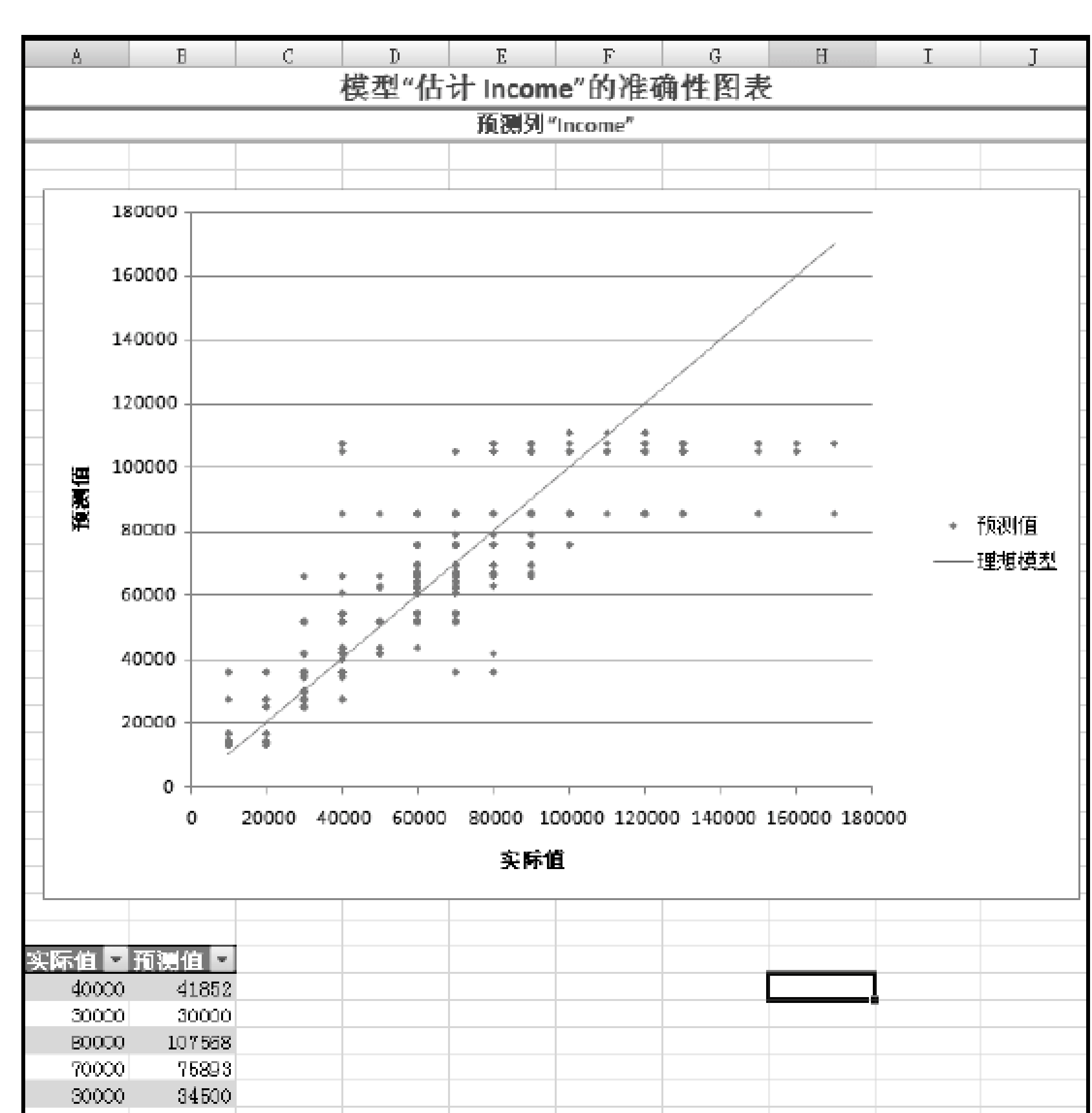


图 6-33　估计模型

6.6.2　分类矩阵

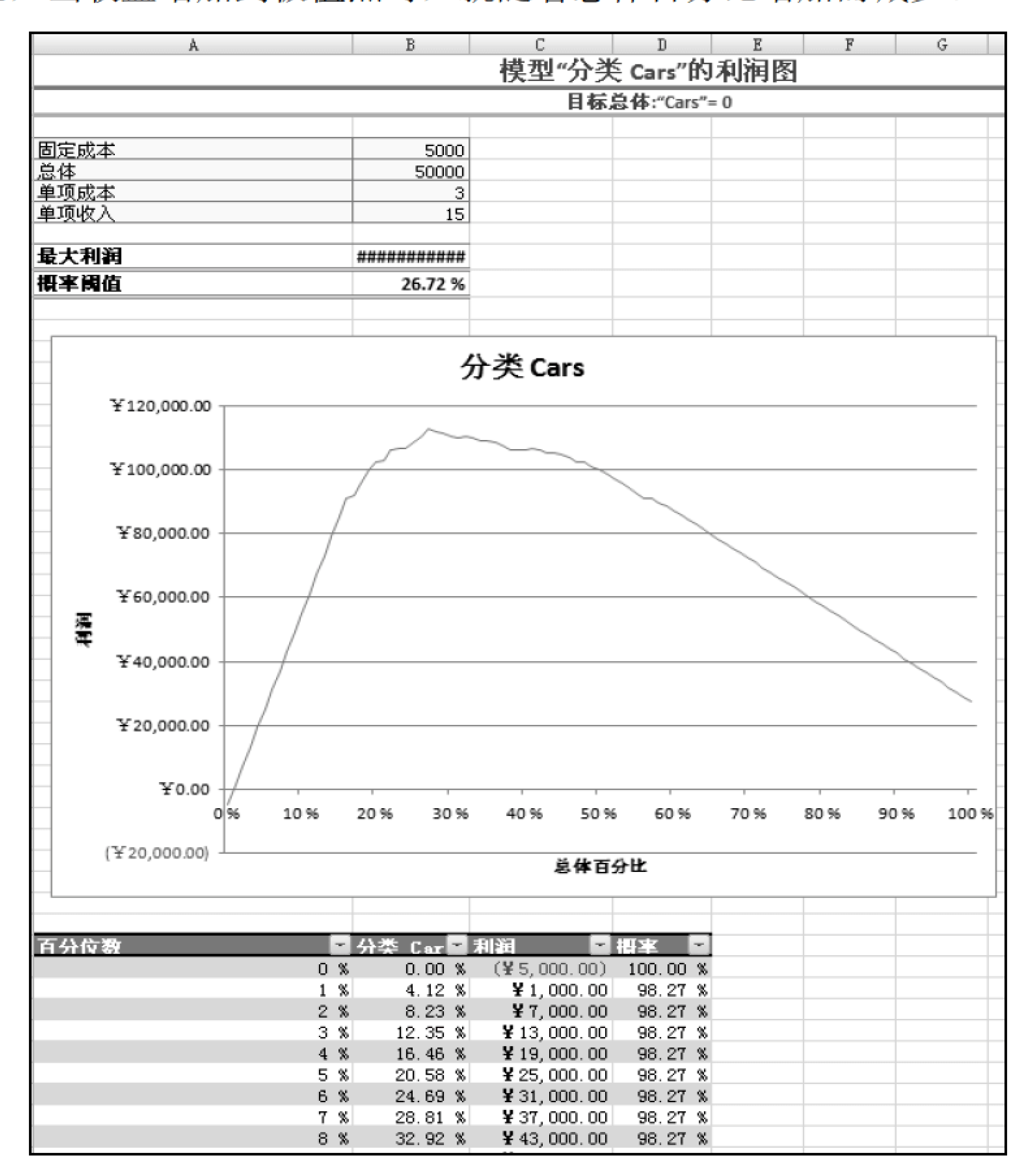
从原始数据中分离的测试集可以用于测试模型的预测效果。通过比较测试数据中的实际值与模型预测结果，可以建立分类矩阵，如图 6-34 所示。

A	B	C	D	E	F
模型"分类 Cars"对列"Cars"的正确/错误分类的计数					
行对应于预测值					
正确总计:	**70.00 %**	**700**			
错误分类总计:	**30.00 %**	**300**			
百分比结果					
	0(实际)	1(实际)	2(实际)	3(实际)	4(实际)
0	**75.31 %**	5.99 %	1.45 %	0.00 %	0.00 %
1	18.93 %	**82.02 %**	21.45 %	38.82 %	33.33 %
2	5.76 %	8.61 %	**73.91 %**	17.65 %	8.33 %
3	0.00 %	0.00 %	2.32 %	**10.59 %**	1.67 %
4	0.00 %	3.37 %	0.87 %	32.94 %	**56.67 %**
正确	**75.31 %**	**82.02 %**	**73.91 %**	**10.59 %**	**56.67 %**
分类错误	**24.69 %**	**17.98 %**	**26.09 %**	**89.41 %**	**43.33 %**
计数结果					
	0(实际)	1(实际)	2(实际)	3(实际)	4(实际)
0	**183**	16	5	0	0
1	46	**219**	74	33	20
2	14	23	**255**	15	5
3	0	0	8	**9**	1
4	0	9	3	28	**34**
正确	**183**	**219**	**255**	**9**	**34**
分类错误	**60**	**48**	**90**	**76**	**26**

图 6-34　分类矩阵

6.6.3　利润图

针对分类模型建立利润图，如图 6-35 所示。图形中的纵轴代表收益，横轴代表总体中数据的百分比，当收益增加到极值点时，就随着总体百分比增加而减少。



固定成本	5000
总体	50000
单项成本	3
单项收入	15
最大利润	###########
概率阈值	26.72 %

百分位数	分类 Car	利润	概率
0 %	0.00 %	(¥5,000.00)	100.00 %
1 %	4.12 %	¥1,000.00	98.27 %
2 %	8.23 %	¥7,000.00	98.27 %
3 %	12.35 %	¥13,000.00	98.27 %
4 %	16.46 %	¥19,000.00	98.27 %
5 %	20.58 %	¥25,000.00	98.27 %
6 %	24.69 %	¥31,000.00	98.27 %
7 %	28.81 %	¥37,000.00	98.27 %
8 %	32.92 %	¥43,000.00	98.27 %

图 6-35　利润图

6.7　模 型 用 法

浏览功能和查询功能用于浏览或查询现有的数据挖掘模型。浏览功能非常好用，可以将构建好的利润图复制到 Excel 上，而且非常美观；可以从利润图中找一个重要案例，将该案例的详细数据提取到 Excel 上。不论利润图或详细数据表，都能够复制到 Excel 上，这对分析人员作报告或其他分析都非常方便好用。

6.7.1　浏览功能

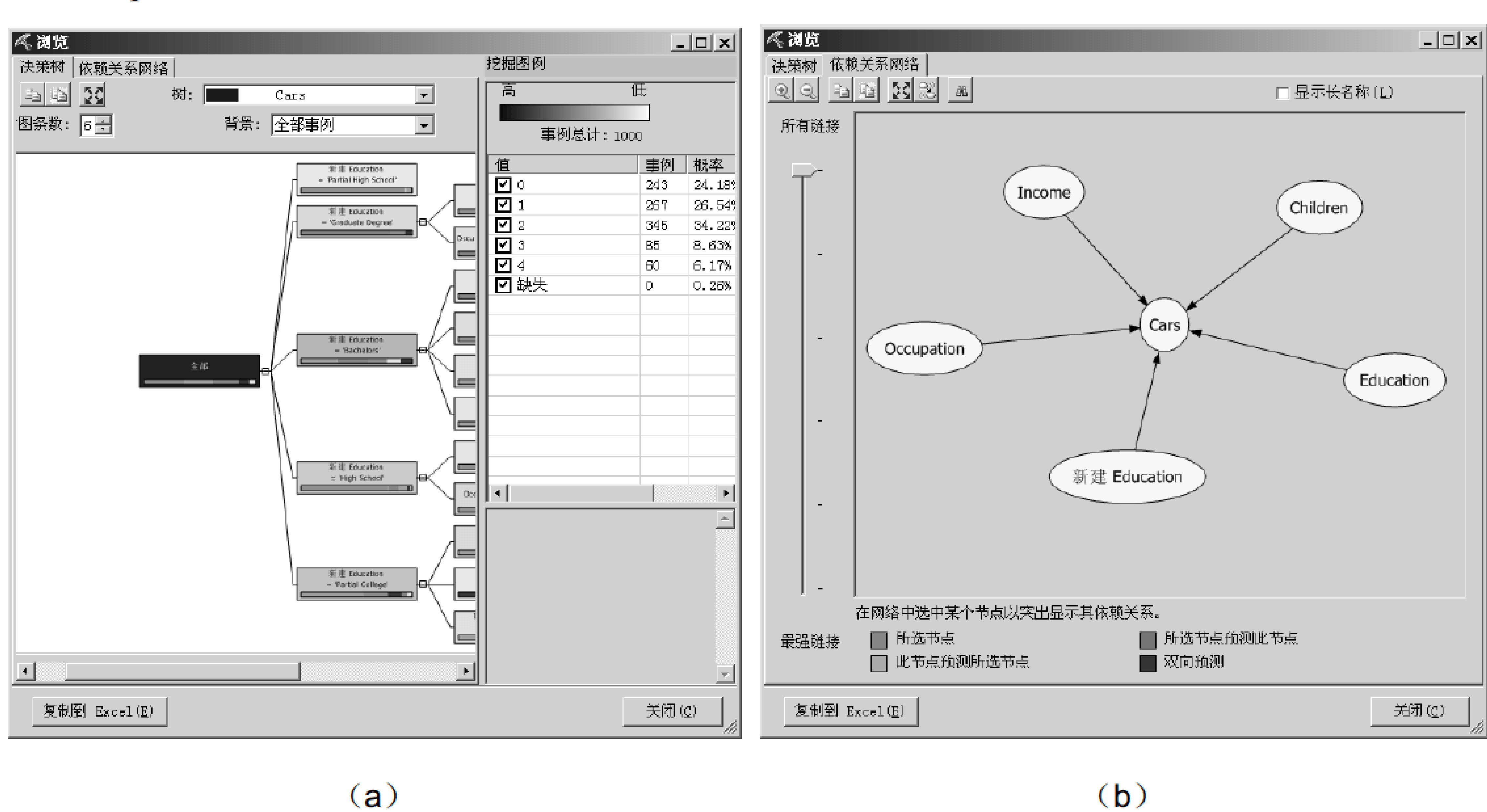
浏览功能的用法如下：

Step1：选择现有模型，如图 6-36 所示。



图 6-36　选择模型

Step2：浏览模型，如图 6-37 所示。



（a）　　　　　　　　　　　　（b）

图 6-37　浏览模型

Step3：复制到 Excel，如图 6-38 所示。

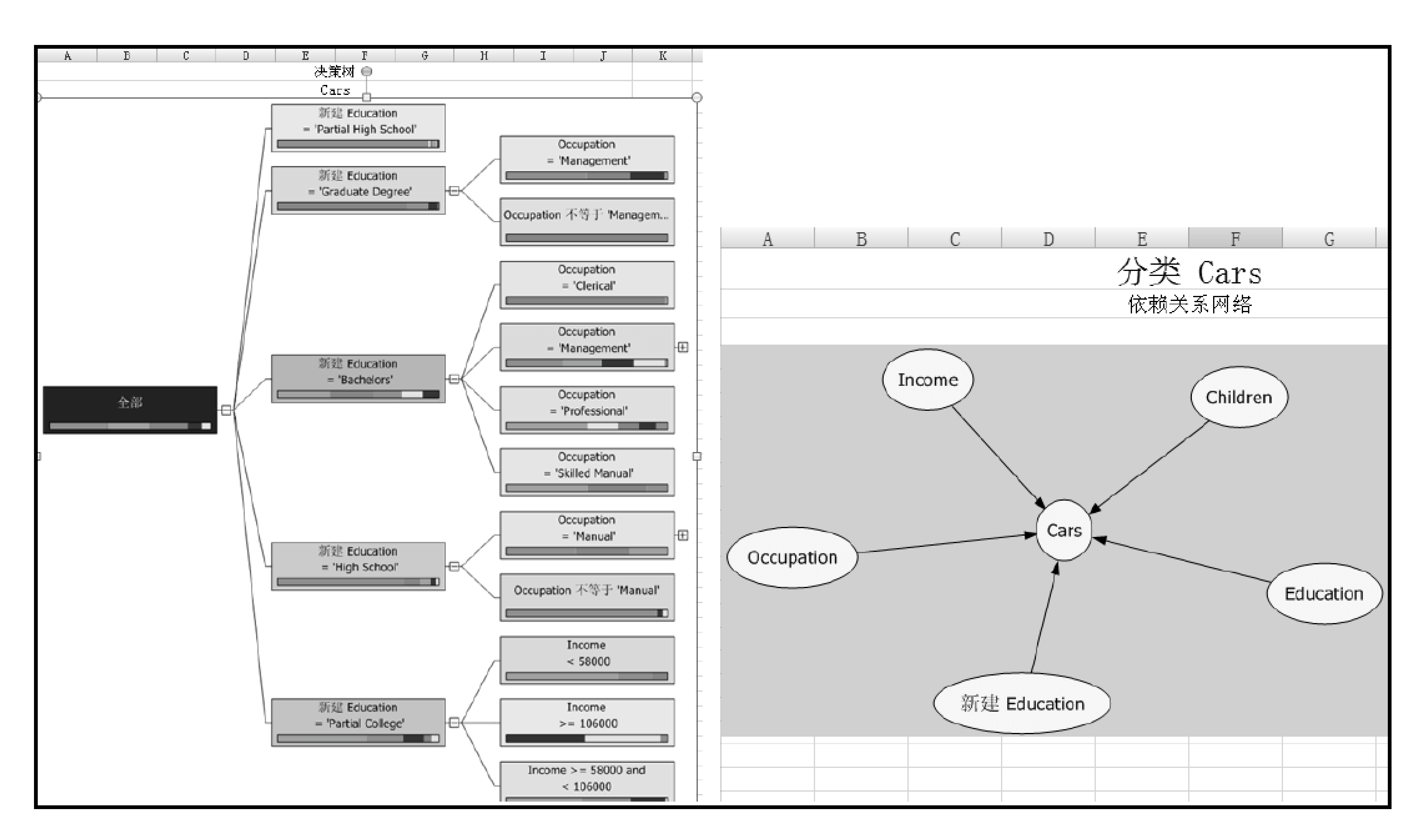


图 6-38 复制到 Excel

Step4：钻取。

（1）在 Education=“High School”图上右击，在弹出的快捷菜单中选择【钻取】命令，如图 6-39 所示。

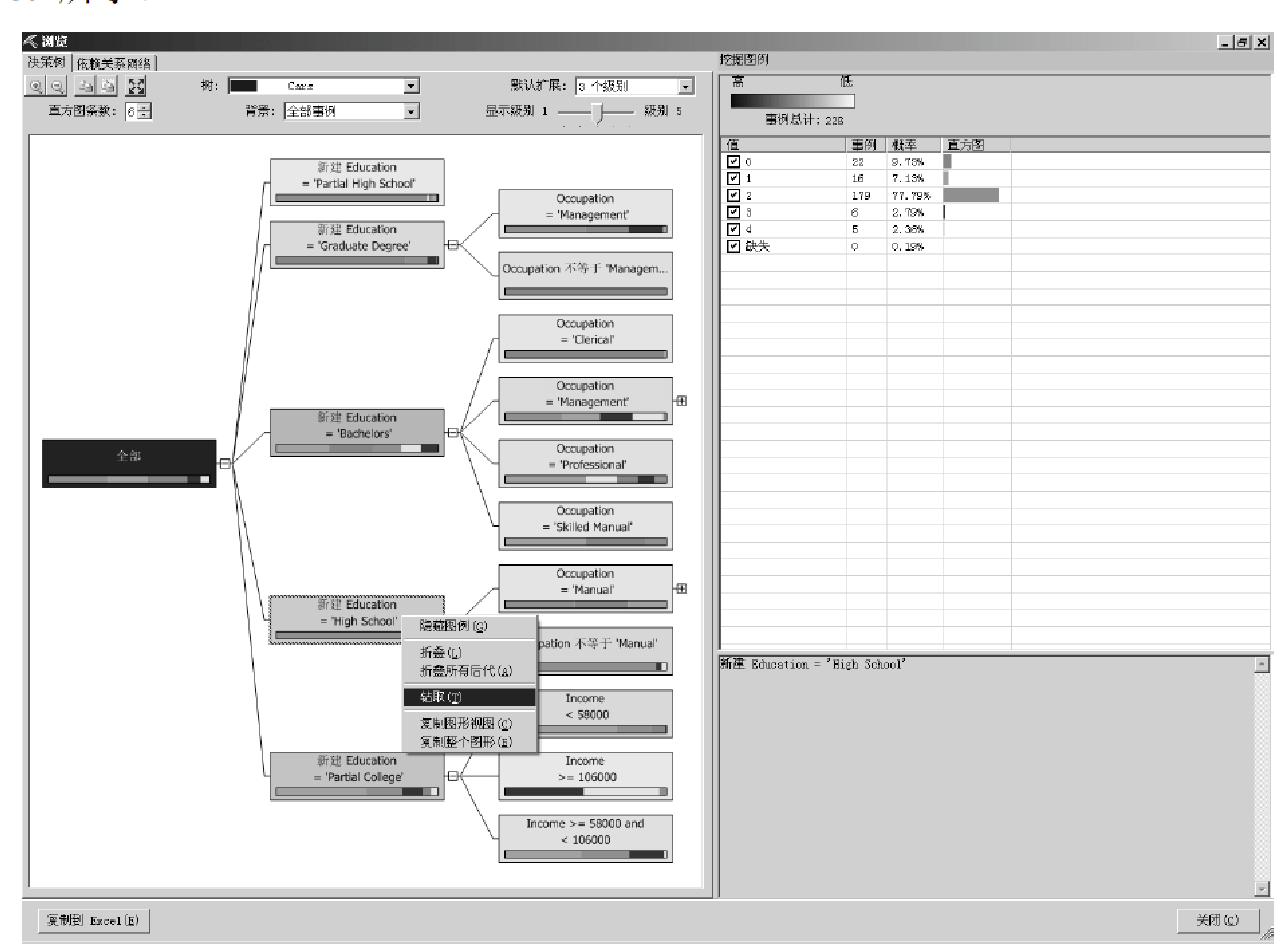


图 6-39 选择【钻取】命令

（2）依据该案例条件的所有数据存在一个新的工作表中，如图 6-40 所示。

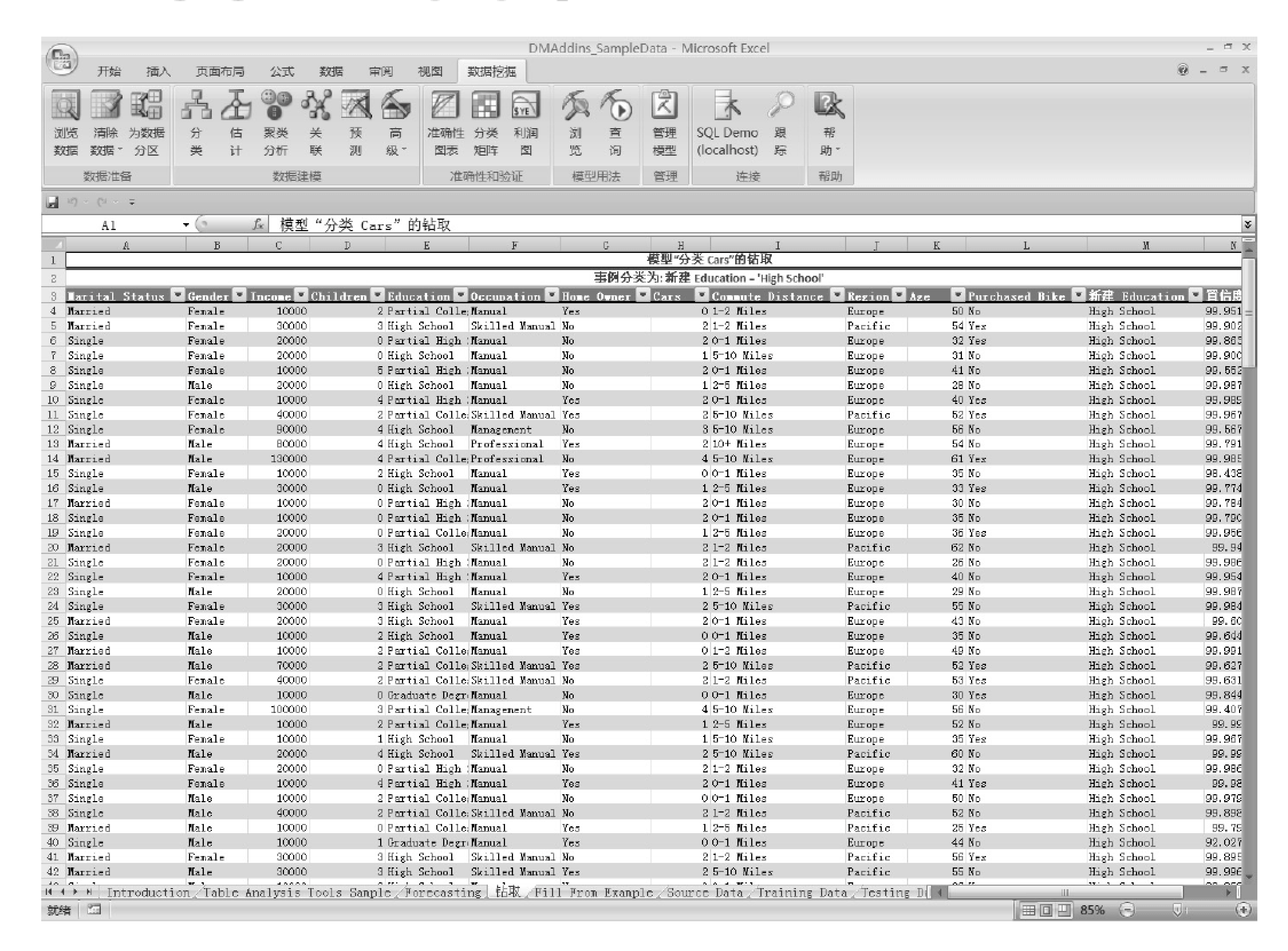


图 6-40　新的工作表

6.7.2　查询功能

对现有的模型建立数据挖掘进行预测查询，可以通过数据挖掘高级查询编辑器撰写 DMX 查询语言，方便查询预测，如图 6-41 所示。

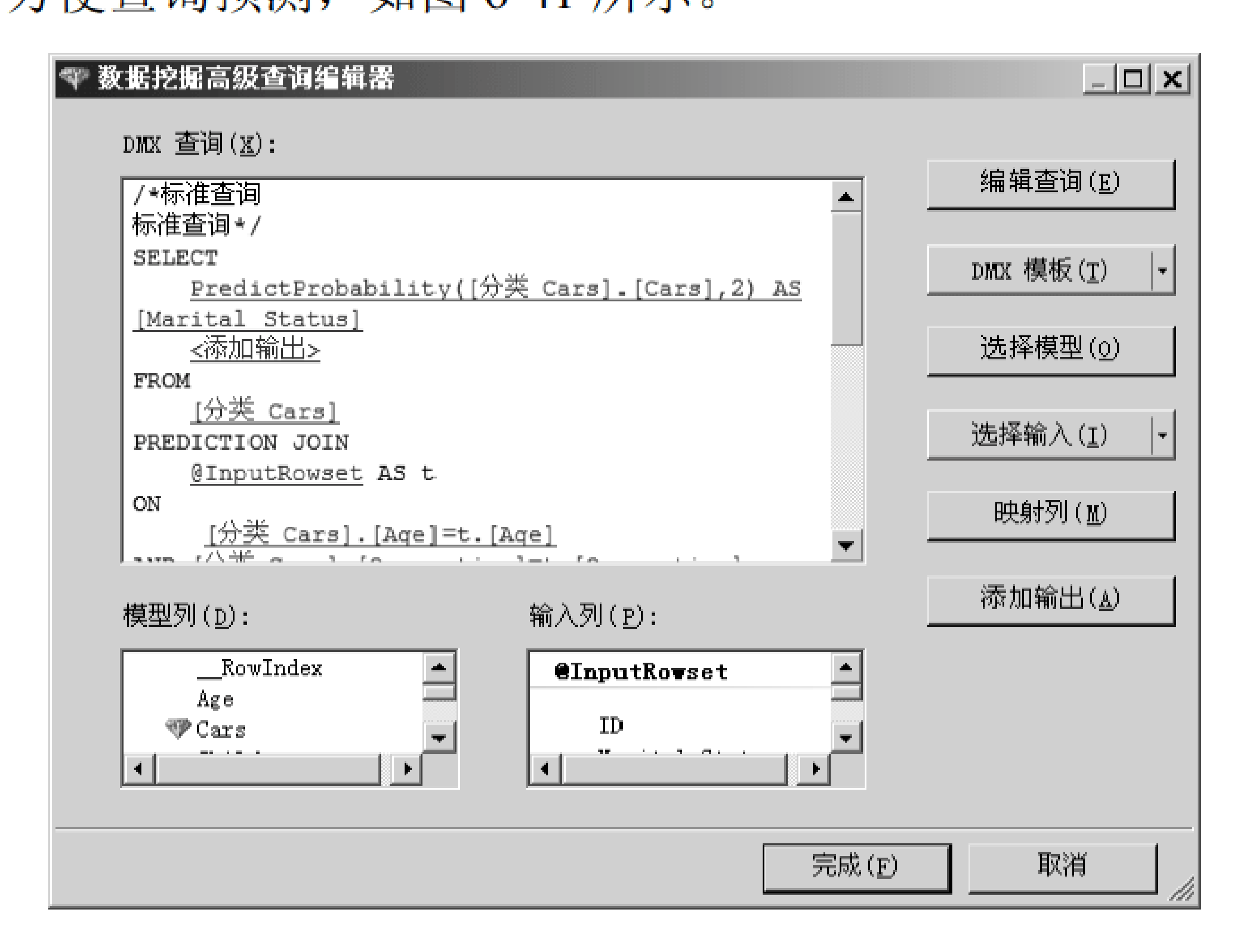


图 6-41　数据挖掘高级查询编辑器

6.8 模 型 管 理

要管理模型，当然是先建立至少一个模型后，才能再进行管理。此功能可以对已经建立的模型进行更名、删除、清除、重新处理、导出、导入等动作，如图 6-42 所示。

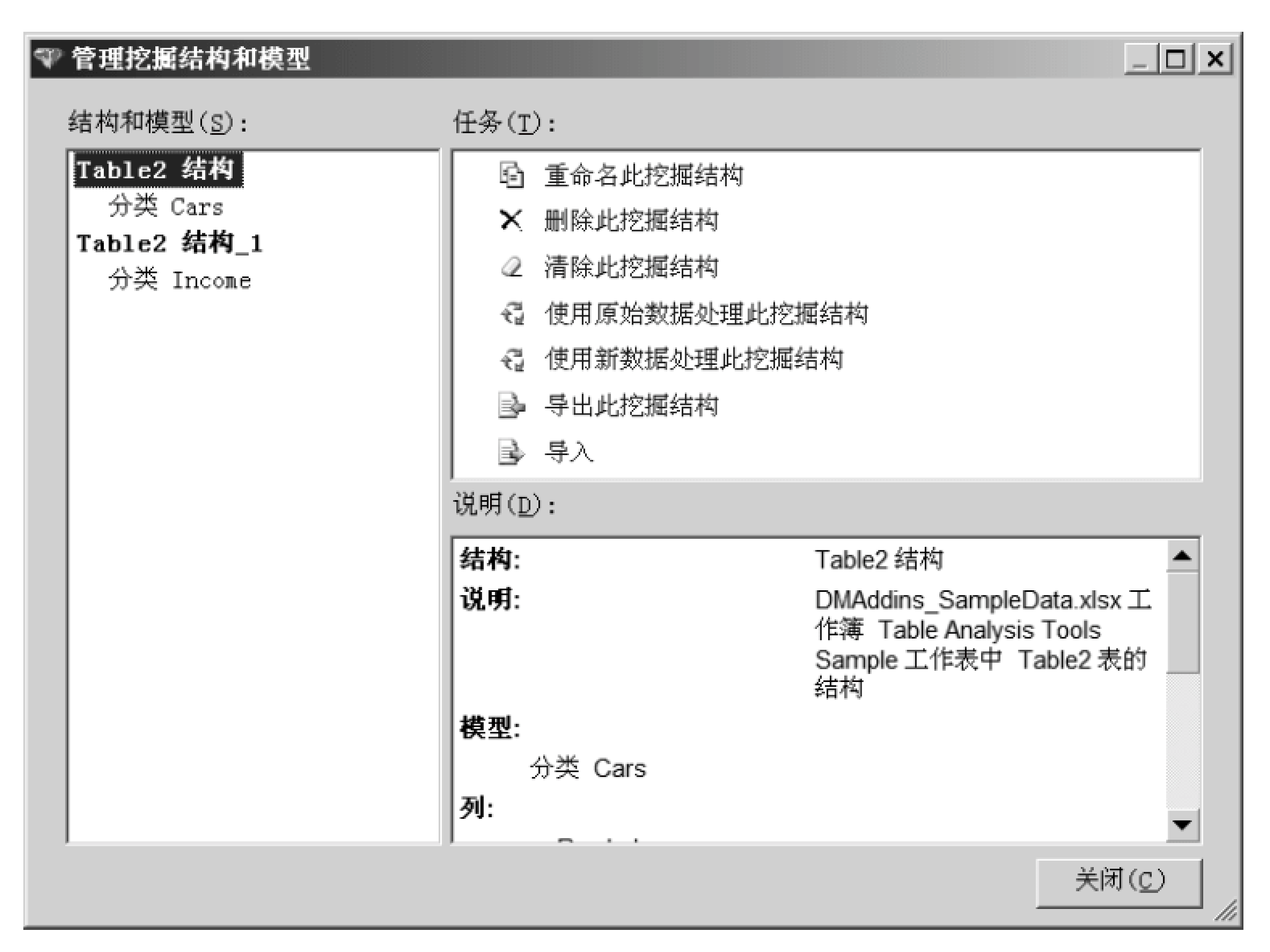


图 6-42 管理挖掘结构和模型

6.8.1 重新命名挖掘模型

单击 重命名此挖掘结构 ，可重新命名挖掘模型。

输入新的挖掘结构名称，如图 6-43 所示。



图 6-43 输入新名称

6.8.2 删除挖掘结构

单击 删除此挖掘结构 ，可删除挖掘结构。

删除之前会询问确认是否删除，如图 6-44 所示。



图 6-44　确认删除

6.8.3　清除挖掘结构

单击 清除此挖掘结构 ，可清除挖掘结构。

清除之前会询问确认是否清除，如图 6-45 所示。



图 6-45　确认清除

6.8.4　用原始数据处理挖掘结构

单击 使用原始数据处理此挖掘结构 ，可用原始数据处理挖掘结构。

重新处理挖掘结构之前，会再确认是否重新处理此模型，如图 6-46 所示。

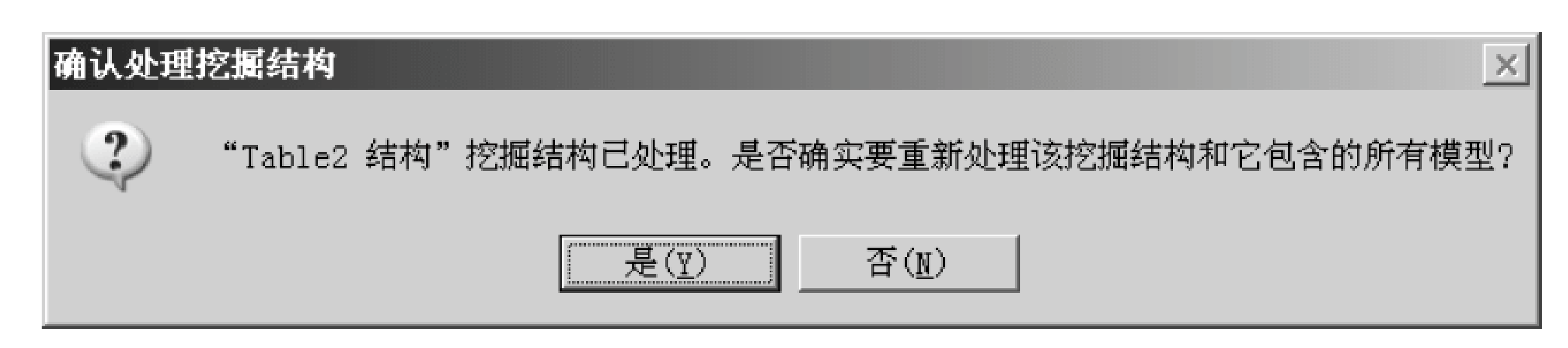


图 6-46　确认处理

6.8.5　用新数据处理挖掘结构

单击 使用新数据处理此挖掘结构 ，可用新数据处理挖掘结构。

重新处理挖掘结构前，会再确认是否重新处理此模型，单击【是】按钮后，选取重新处理挖掘结构的数据来源，单击【下一步】按钮，如图 6-47 所示。

重新设定挖掘结构的输入与输出之间的数据列对应，单击【完成】按钮，如图 6-48 所示，数据挖掘结构就会重新处理。



图 6-47　选择源数据



图 6-48　指定列映射

6.8.6　导出挖掘结构

单击 导出此挖掘结构，可导出挖掘结构。

输入导出的文件名及位置，如图 6-49 所示。

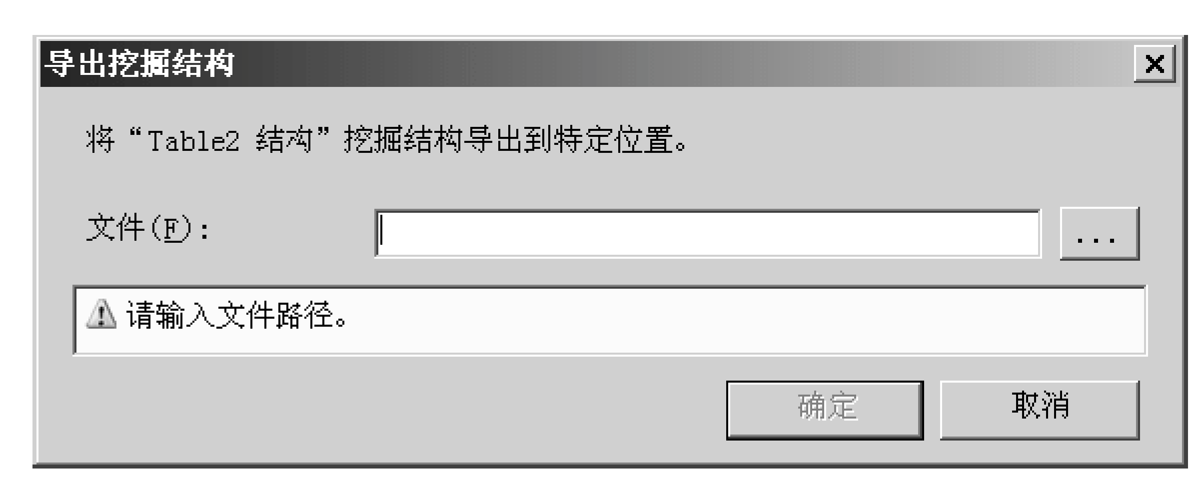


图 6-49　输入文件名及位置

6.8.7 导入挖掘结构

单击 导入 ，可导入挖掘结构。

输入导入的文件名及位置，如图 6-50 所示。

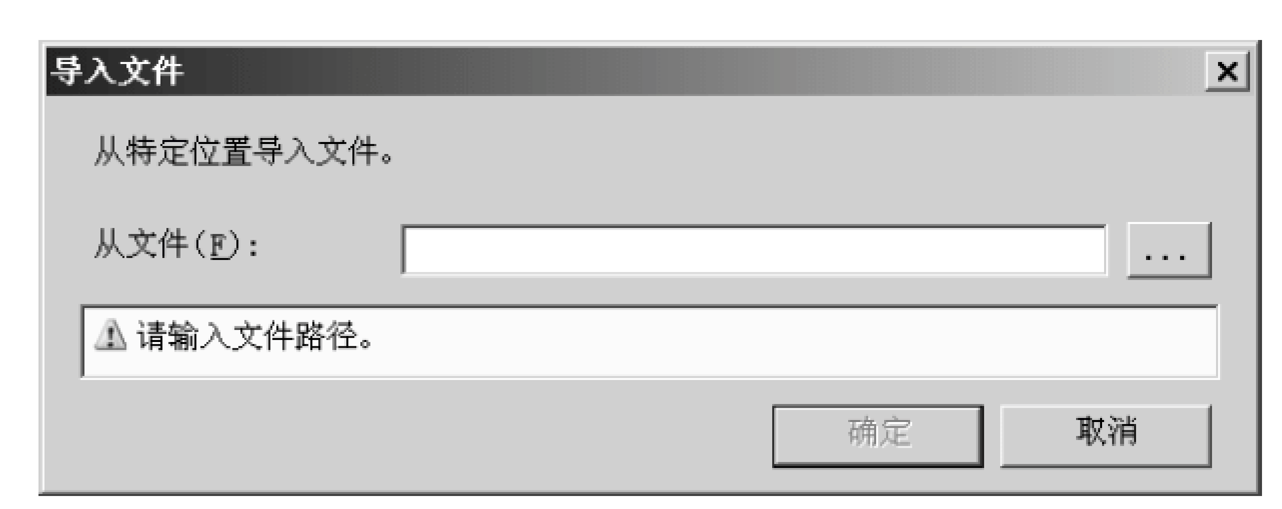


图 6-50 输入文件名及位置

第 7 章　决　策　树

7.1　基 本 概 念

决策树是数据挖掘中的一项主要方法，其方法原理是利用多个预测变量对定性变量进行预测。决策树和判别分析一样，都可以完成分类任务。但是决策树的弹性，使得数据本身更具吸引力。

7.2　决策树模块的建立

决策树模块的建立包括三种类型：针对类别型的预测变量，计算以单变量分裂为基础的二分叉决策树；针对有序型的预测变量，计算以单变量分裂为基础二分叉决策树；针对类别和有序混合型的预测变量，计算以单变量分裂为基础的二分叉决策树。当然也可以建立线性组合的分裂函数（linear combination split），对个体进行划分。

7.3　决策树与判别函数比较

决策树与判别函数的比较，如表 7-1 所示。

表 7-1　决策树与判别函数比较

决　策　树	判 别 函 数
使用系数和判别函数，突出层次性，分类更加准确	使用系数和判别函数，但没有层次性，因而分类效果较差
决策树的预测变量和分类变量之间，可以分别进行独立的回归分析	预测变量与因变量间的关系可视为一个多元回归方程
递归层次结构作为分类原则	利用个体的属性变量的相似性作为判别依据
可以逐个考查决策树的预测变量的影响	所有预测变量同时出现在判别函数中，无法判断每个预测变量的重要性
可处理类别变量、连续变量或混合型的预测变量	一般要求预测变量至少是有序的定性数据

7.4 计 算 方 法

7.4.1 确定预测精度的标准

决策树分析的目的是把个体归入其最有可能正确的类别，因而预测精度的定义就显得相当重要。一般来说，成本指个体是否有发生错误分类的现象，且占所有个体数的比例。进一步讲，成本还可以定义为预测时可以承受的损失。因此，成本越小，混合分类的情形就越小，预测精度越高。

在决策树中，还要考虑个体所属类别的先验概率。如果各个类别的比例相似，或者各个分类中的个体数接近相等，那么可选择相同先验概率；如果各个类别的比例相差较大，可以把样本中的类别比例作为先验概率；如果针对某个类别有着特定的意义或特别的考虑，则可以分别设定不同的先验概率。

7.4.2 选择分裂（分层）技术

各种分裂（分层）技术如表 7-2 所示。

表 7-2　分裂（分层）技术

分裂（分层）技术	说　明
单变量分裂函数	第一个步骤为针对现有决策树选择预测变量和该变量的分裂临界值。计算个体与预测变量间的相关性。如果该预测变量是类别变量，则计算卡方检验的 p-value；如果预测变量为数值变量，则以 ANOVA 计算 p-value
线性组合分裂函数	预测变量假设为数值型。这种以连续预测变量计算线性组合的结果与前一种纯粹以类别尺度预测变量的结果类似
CART 方式的完全搜寻（单变量）	在分类树模块中，提供三种拟合优度检查的方法：Gini 指标、Chi-Square 法和 G-Square 法

7.4.3 定义停止分裂（分层）的时间点

如果因变量的可观察分类或者分类树分析中的预测变量的层次水平测量错误或存在噪音，就无法得到最终的分类节点。决策树一般提供两个功能选项可以控制停止分裂：（1）最终节点中应该包含的最小的个体数。在分类树执行的过程中，程序会计算落入节点数的个数直至满足这一条件，才会停止。（2）指定个体所属类别的比例。分类过程一直持续到纯的最终节点出现或者没有任何分类超过该比例。如果先验概率相同，且各分类的个数相同，那么当最终节点为空时，分裂过程自动停止；如果先验概率不等，程序依然会对指定的分类大小与片段数值相比较，直至满足预设条件时才停止。

7.4.4 选择适当大小的决策树

一般而言，决策树的大小是任意的，但应在保证预测精度前提下，省略不必要的分支。在微软的分类树算法中，有多种不同的选取策略可选择使用，如表 7-3 所示。

表 7-3 各种选取策略

策　　略	说　　明
面向实际的交叉验证（fact style direct stopping）	采用 FACT-style direct stopping 的停止规则，诊断现有信息用以定义树状结构大小的合理性；采用交叉确认的方法检查合理性
验证集交叉验证（test sample cross validation）	仅在预留的验证样本中进行交叉验证
K 重交叉验证（k-fold cross validation）	样本分为大小相同的 V 个子样本，每次任意抽取一个子样本作为验证数据集，余下的（V–1）个子样本作为训练集
整体交叉验证（global cross validation）	将全部分析依据制定的次数复制（重叠），并划分部分片段为样本。将此片段样本视为查看样本，与重复的学习样本进行交叉确认
最小成本复杂度交叉验证（minimal cost complexity cross validation pruning）	在分类树模块中，当停止规则为错误分类率时，最小成本复杂度交叉验证较优

7.5 Excel 2007 决策树算法

微软的决策树算法同时支持离散和连续变量的预测。

Step1：单击【高级】按钮，选择【创建挖掘模型】命令，如图 7-1 所示。



图 7-1 创建挖掘模型

Step2：单击【下一步】按钮，如图 7-2 所示。

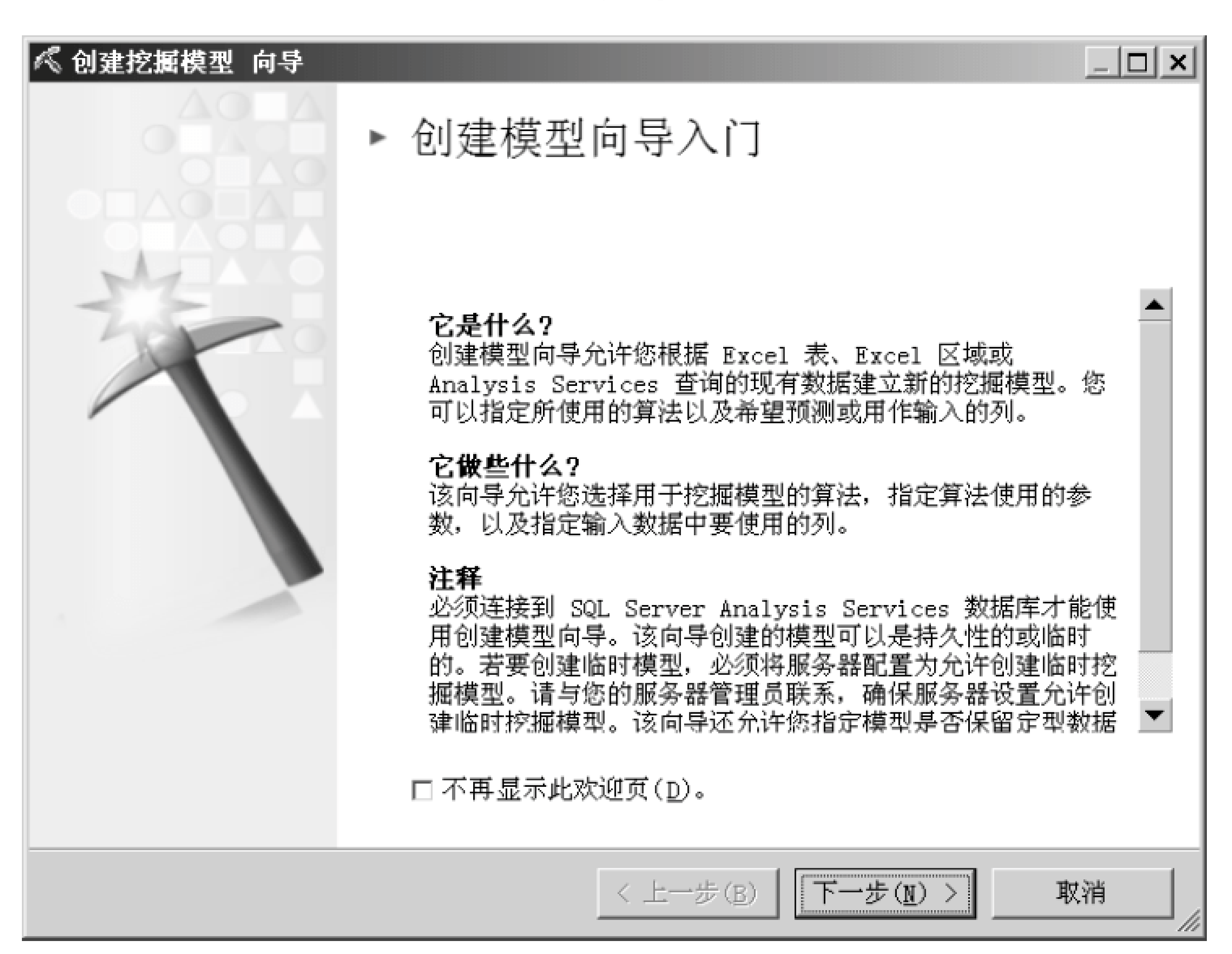


图 7-2　创建模型向导入门

Step3：选择数据表，单击【下一步】按钮，如图 7-3 所示。

图 7-3　选择数据表

Step4：选择挖掘算法，如在下拉列表框中选择 Microsoft 决策树选项，单击【下一步】按钮，如图 7-4 所示。

Step5：变量选择，设 Income 为因变量，并设为【仅预测】，单击【下一步】按钮，如图 7-5 所示。



图 7-4 选择挖掘算法



图 7-5 选择列

Step6：单击【完成】按钮完成设置，如图 7-6 所示。

Step7：决策树展开。由图 7-7 所示的决策树展开可知，当 Occupation 是 Professional，下一个重要变量为 Region；当 Occupation 是 Skilled Manual，下一个重要变量为 Education；当 Occupation 是 Management，下一个重要变量为 Age。

Step8：依赖关系网络，可按照关系的强弱判别 Income 和其他解释变量的关系大小。其中所选与因变量 Income 关系相关的变量包括 Region、Age、Education、Children、Cars 和 Occupation，如图 7-8 所示。



图 7-6　完成

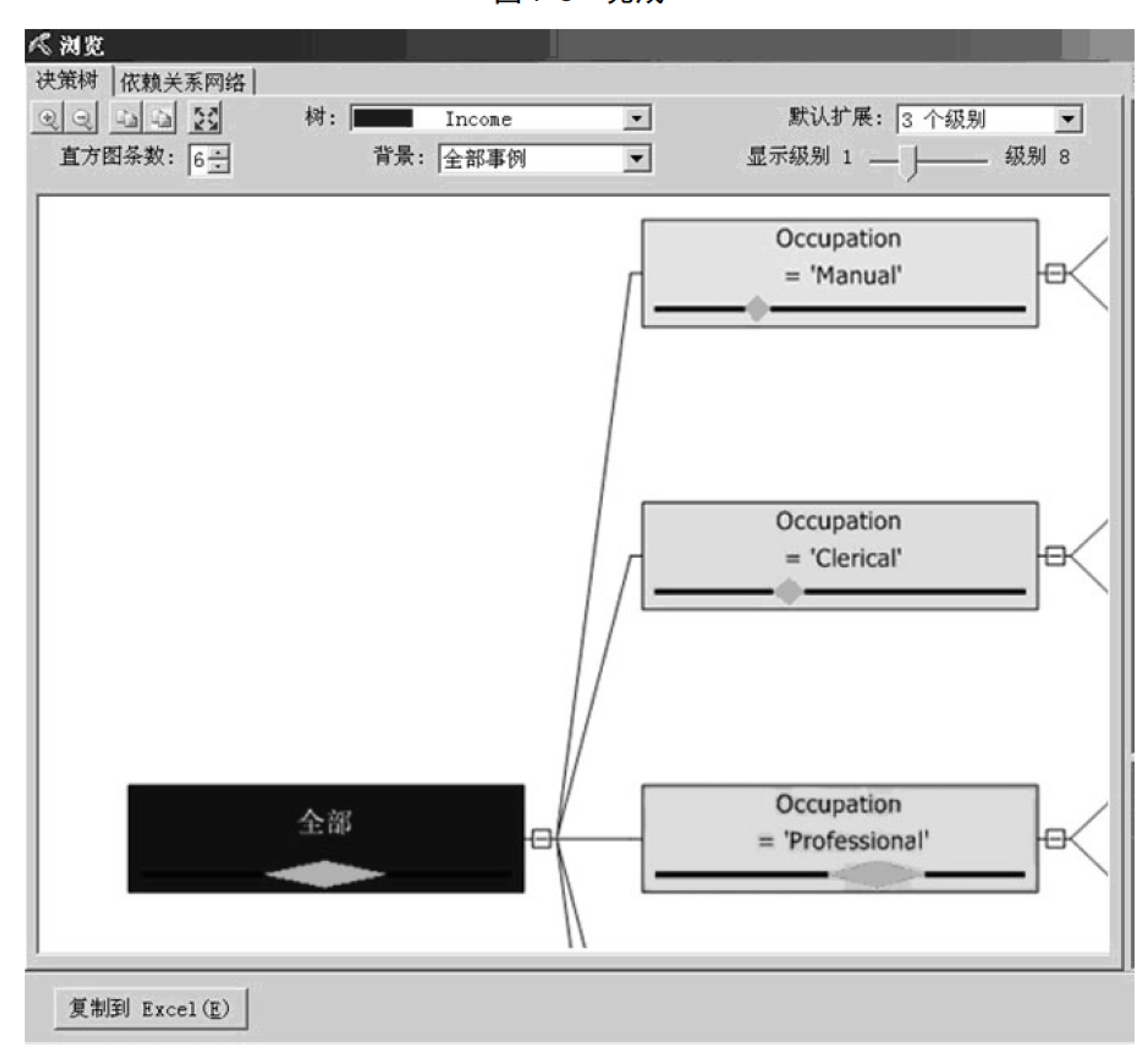


图 7-7　决策树展开

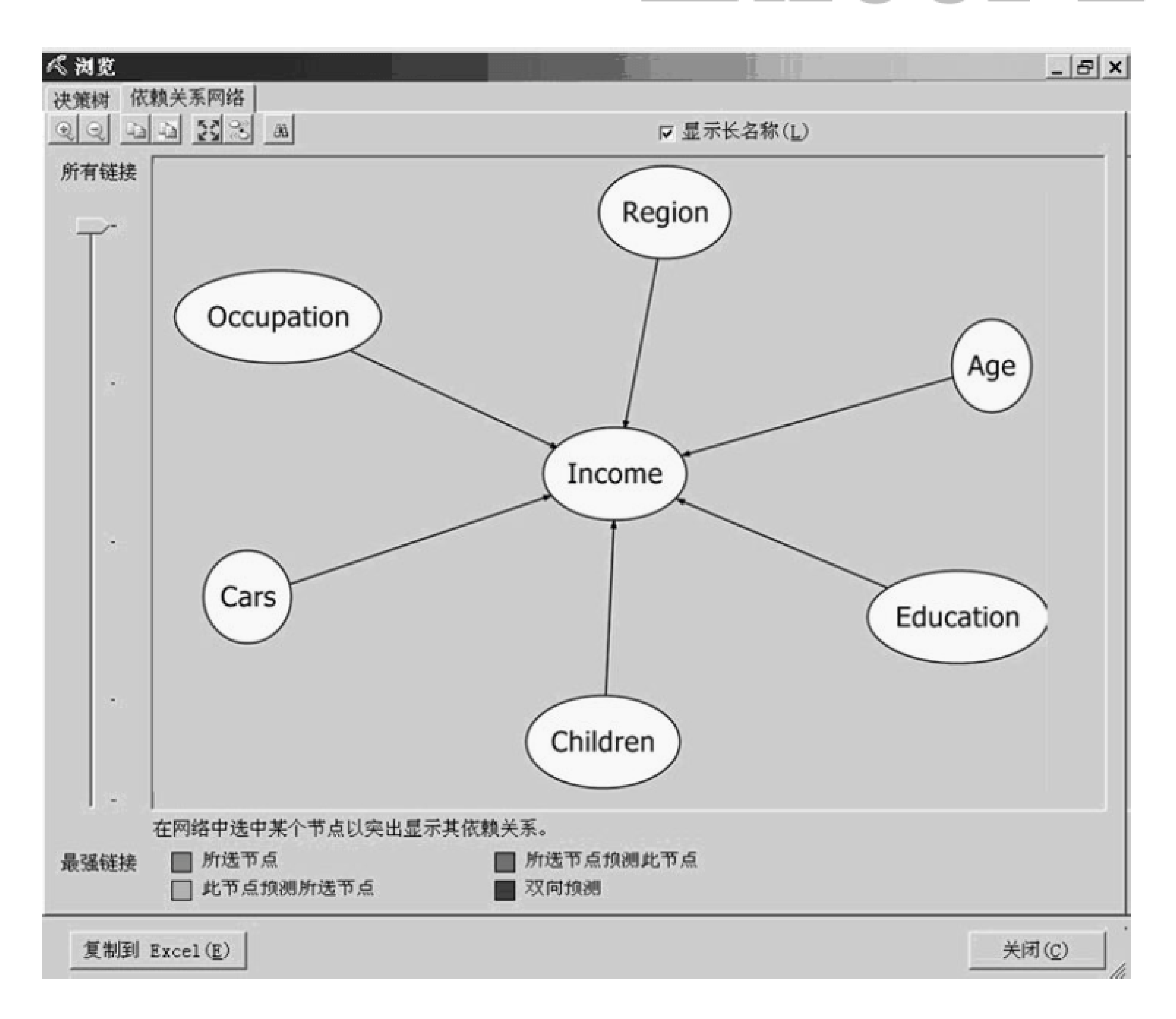


图 7-8　与因变量相关的变量

Step9：单击【复制到 Excel】按钮，将依赖关系网络复制到 Excel 文件中，如图 7-9 所示。

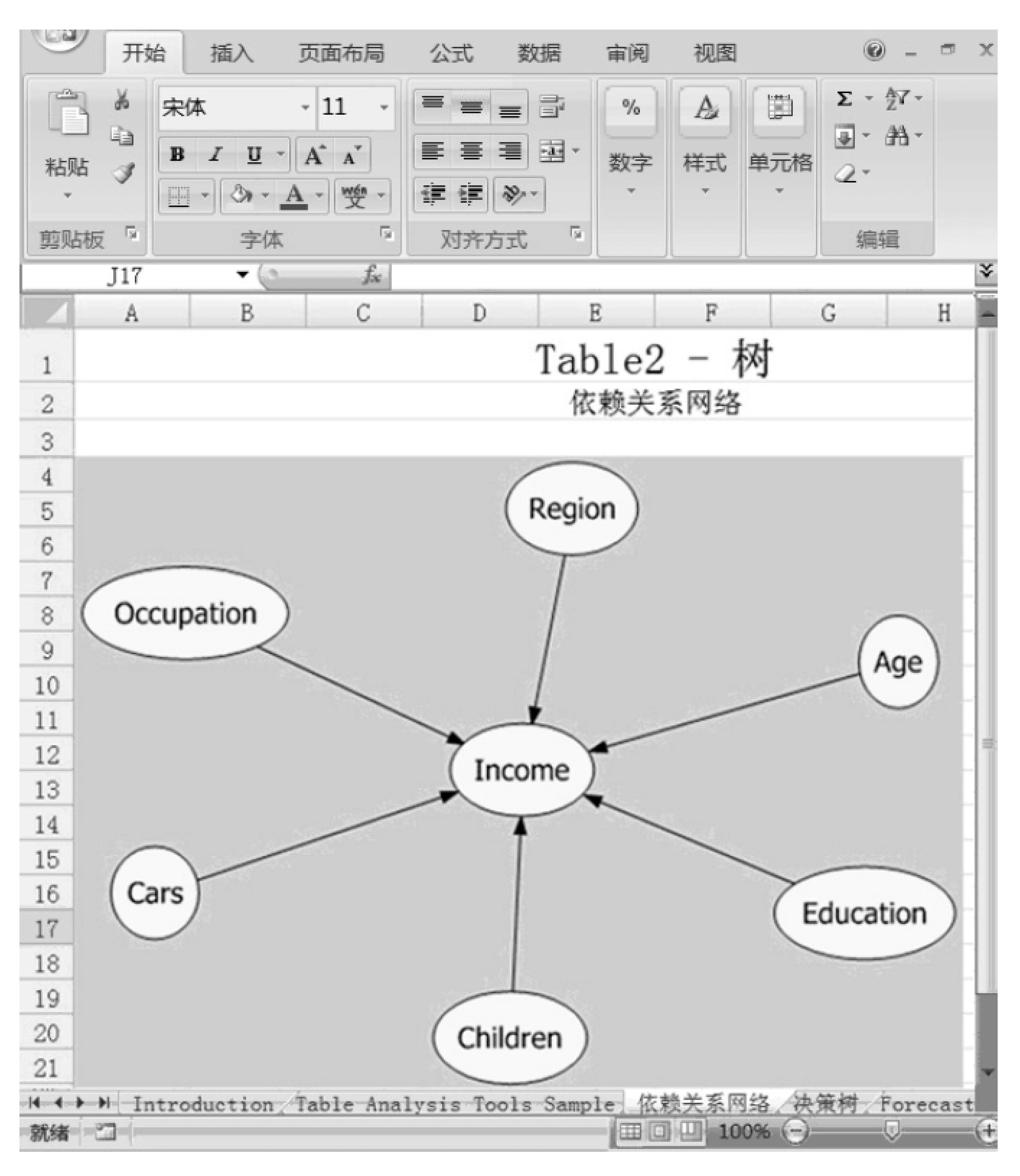


图 7-9　复制到 Excel

Step10：单击【精确度图表】按钮，弹出的图 7-10 所示的【准确性图表向导入门】窗口。

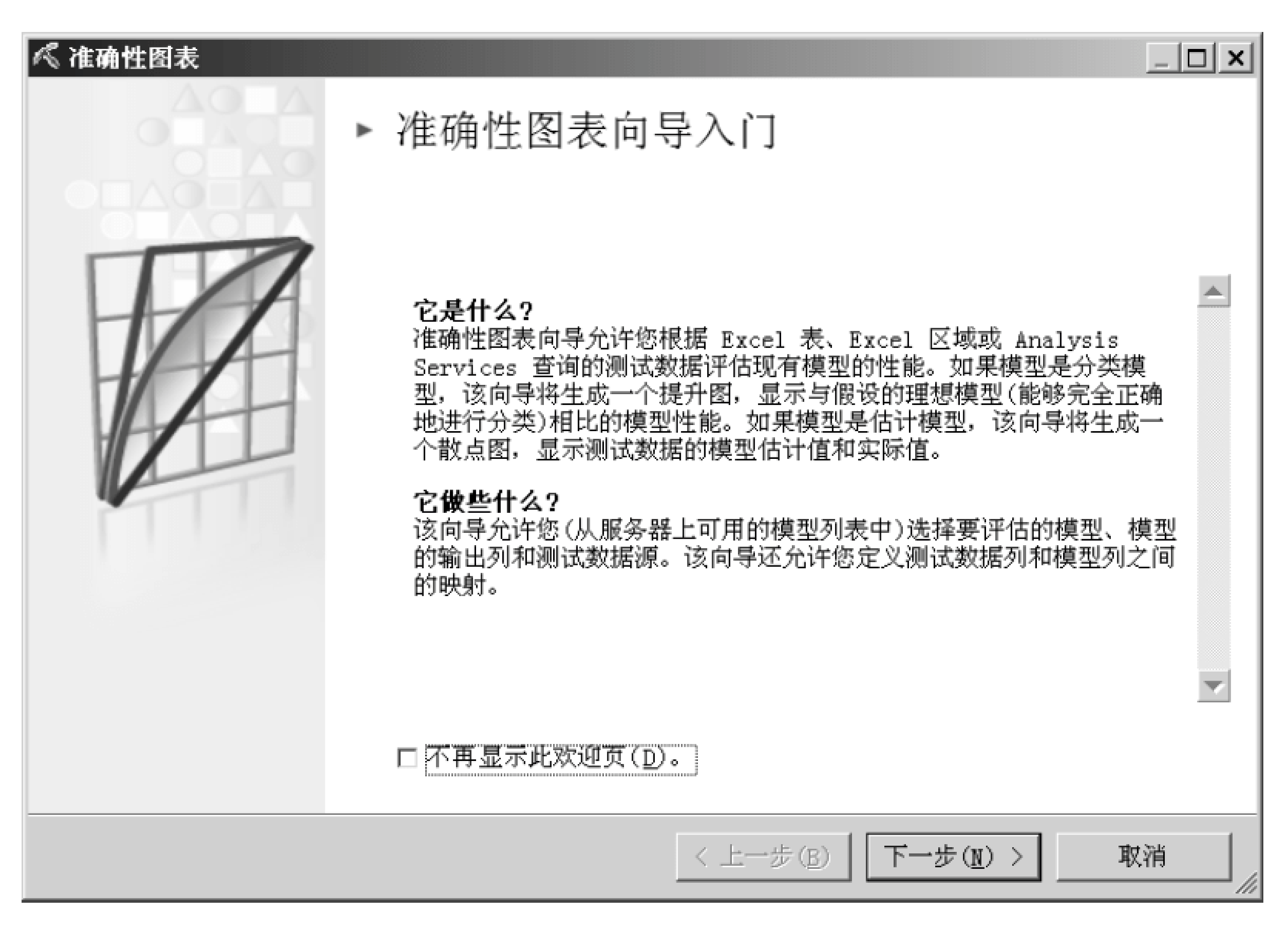


图 7-10 【准确性图表向导入门】窗口

Step11：单击【下一步】按钮，弹出的如图 7-11 所示的【选择模型】窗口。

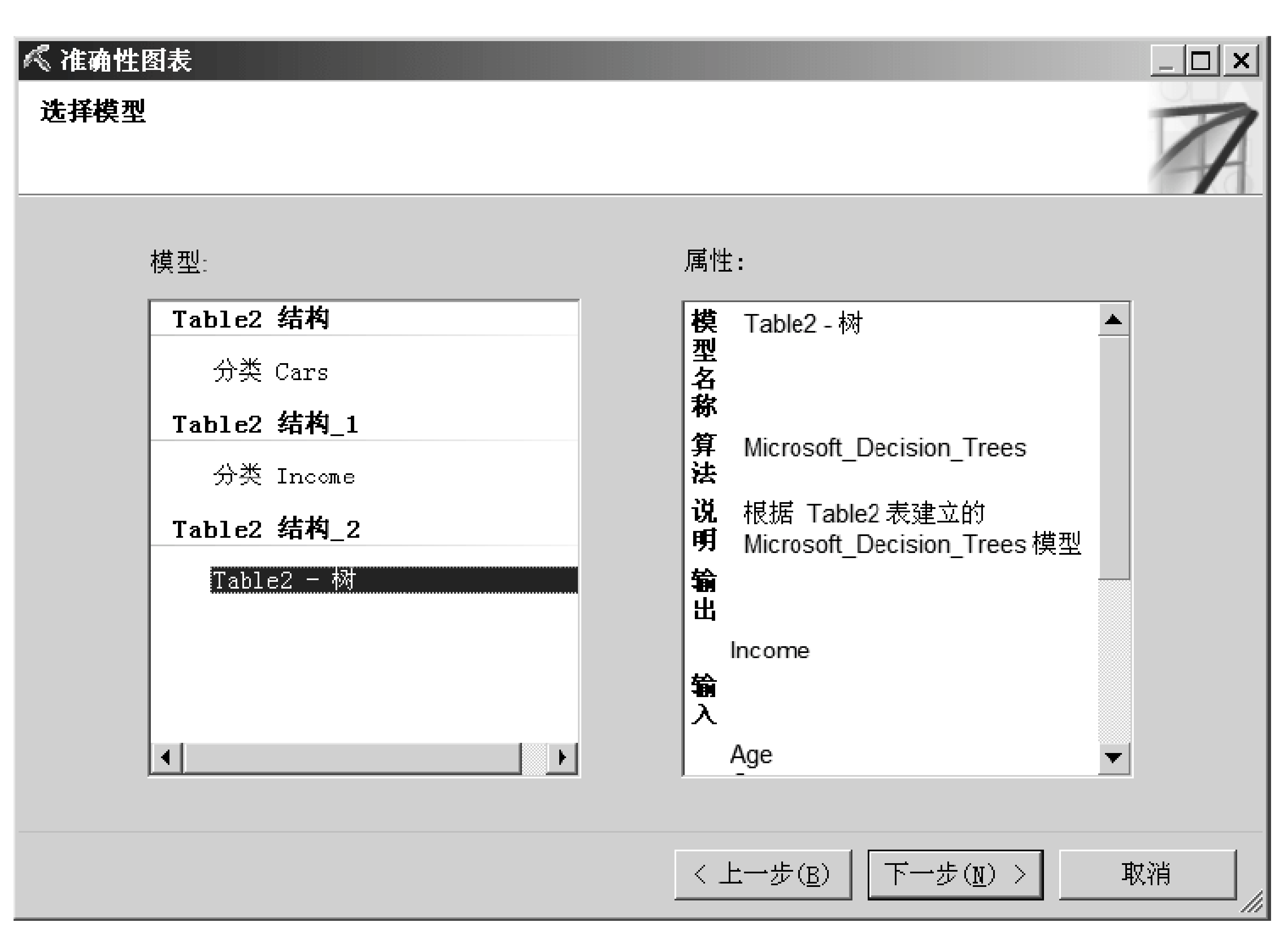


图 7-11 【选择模型】窗口

Step12：选择模型后单击【下一步】按钮，弹出如图 7-12 所示的【指定要预测的列和要预测的值】窗口。

Step13：选择数据表，如图 7-13 所示。

Step14：单击【完成】按钮，如图 7-14 所示。

Step15：显示精确度图，如图 7-15 所示。

Step16：显示精确度表，如图 7-16 所示。

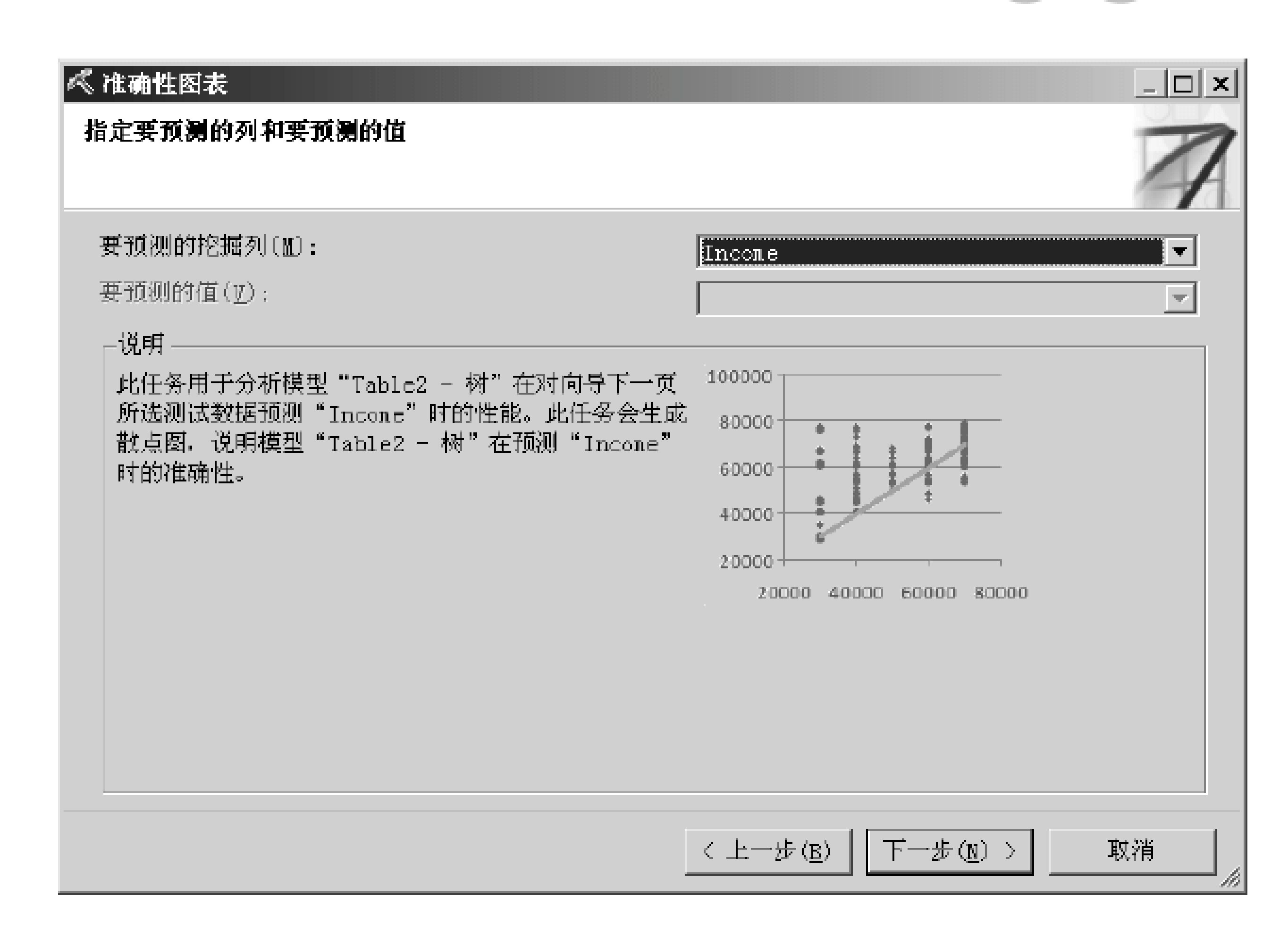


图 7-12 【指定要预测的列和要预测的值】窗口



图 7-13 选择数据表



图 7-14 单击【完成】按钮

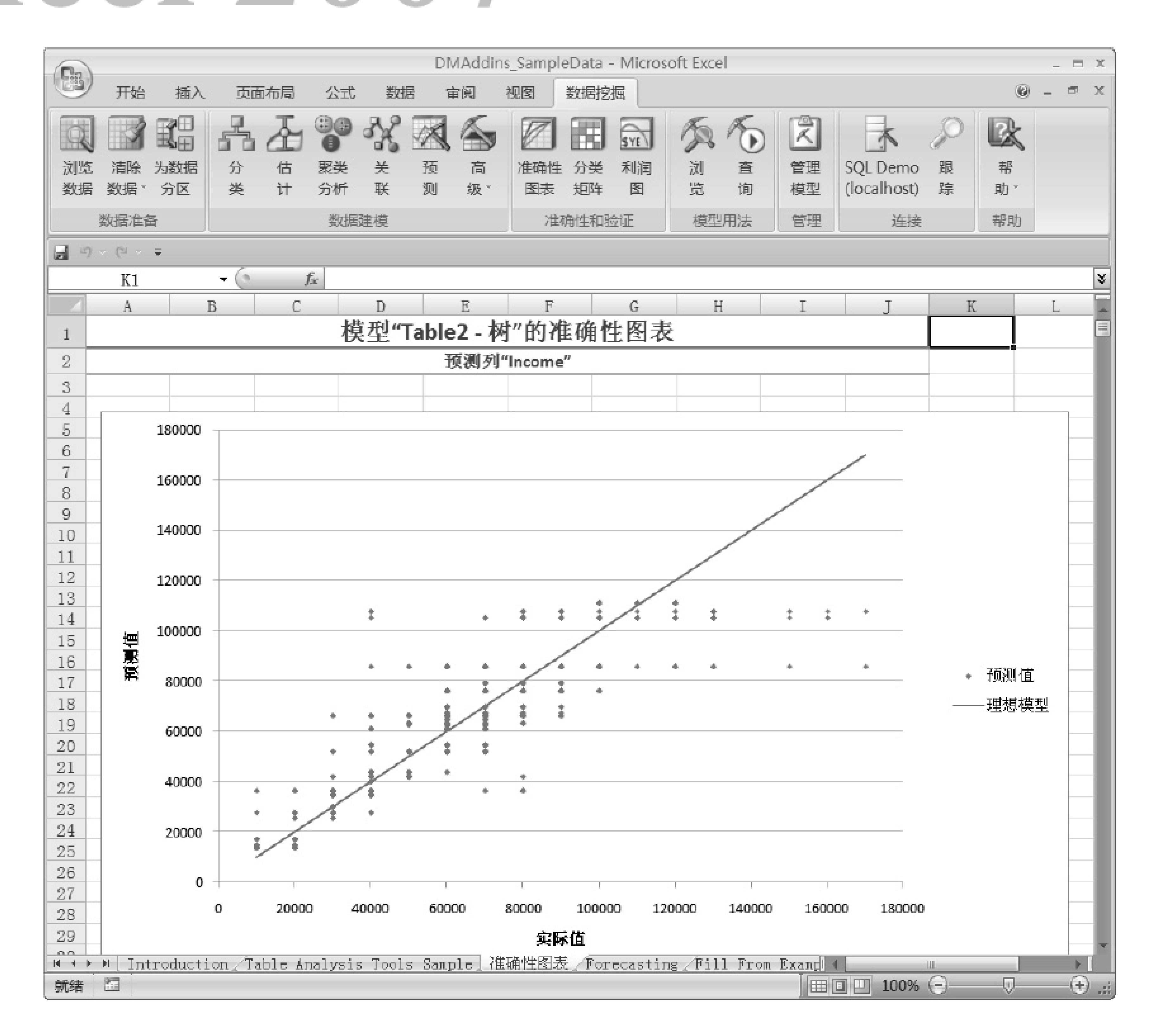
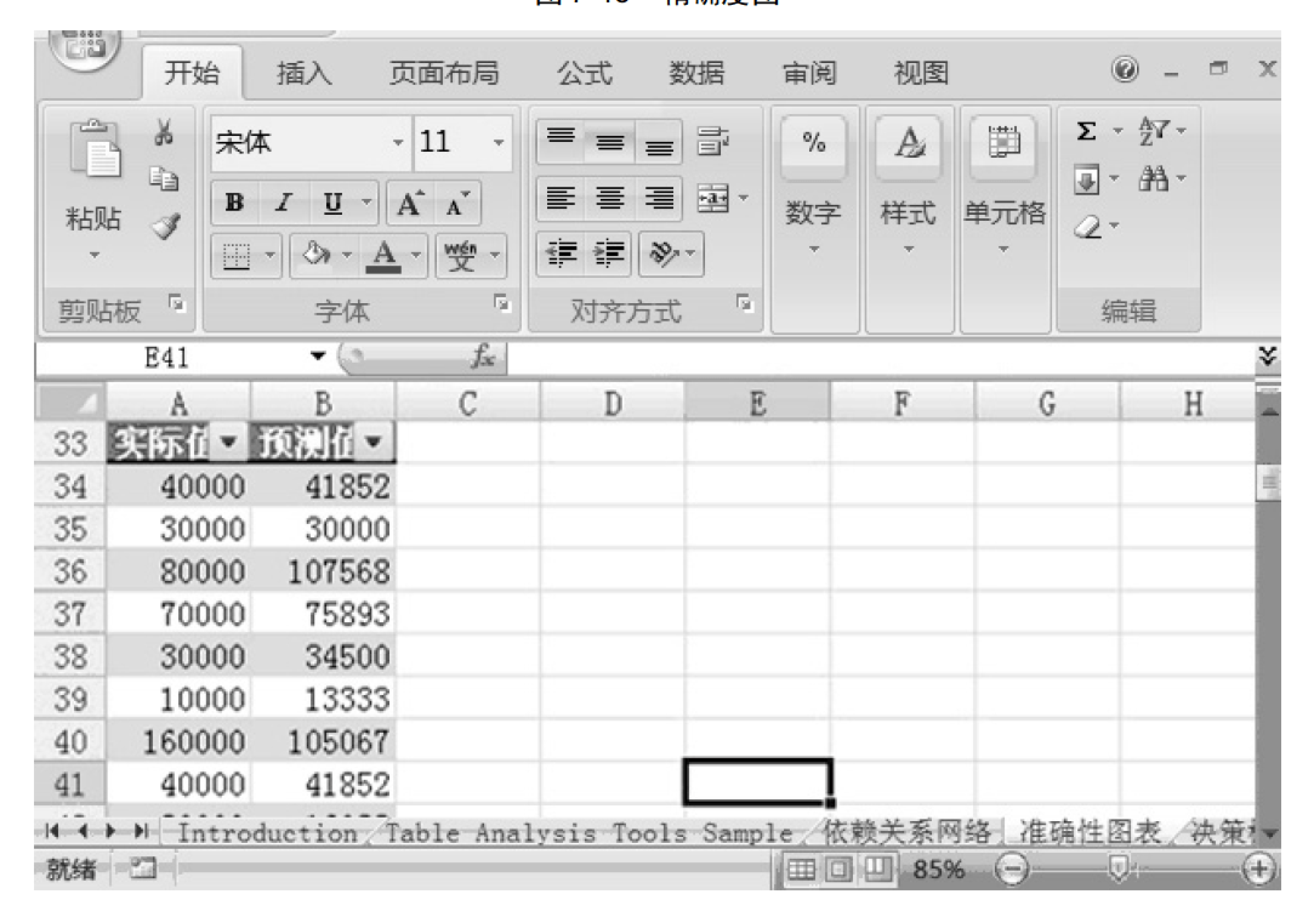


图 7-15　精确度图



	实际值	预测值
34	40000	41852
35	30000	30000
36	80000	107568
37	70000	75893
38	30000	34500
39	10000	13333
40	160000	105067
41	40000	41852

图 7-16　精确度表

第 8 章　贝叶斯概率分类

8.1　基 本 概 念

贝叶斯分类（bayes classifier）是一种简单实用的分类方法。在分类之前，需知道总体中不同类别的比例构成，通过训练样本，学习并产生这些训练样本的分类规则，再用这些分类规则对其他个体进行分类预测。一般而言，分类变量可能出现两种以上不同的值，而目标变量多为二元的相对状态，如“是 / 否”，“好 / 坏”，“对 / 错”或者“上 / 下”等。

简单贝叶斯分类（naive bayes classifier）是根据贝叶斯定理，交换先验概率（prior）和后验概率（posteriori），在分类属性相互独立的假设（conditional independence）下预测分类的情形。其公式如下：

$$\begin{aligned} h_{\mathrm{MAP}} &= \underset{h \in V}{\operatorname{argmax}}\, P(h \mid D) \\ &= \underset{h \in V}{\operatorname{argmax}}\, P(D \mid h) P(h) \end{aligned}$$

其中：

h_{MAP}为最大可能的假说（maximum a posteriori）。

D 为训练样本。

V 为假设空间（hypotheses space）。

$P(D|h)$为训练样本的事前概率，对于假说 h 而言，为一常数。

$P(h)$为假说 h 事前概率（尚未观察训练样本时的概率）。

$P(h|D)$为在训练样本 D 集合下，假说 h 出现的条件概率。

简单贝叶斯分类根据训练样本，对于个体的属性值（$a_1, a_2, a_3,\cdots, a_n$）（假设一共有 n 个学习概念的属性 $A_1, A_2,\cdots, A_n$，a_1 为 A_1 相对应的属性值），指派具有最高概率值的类别（C 表示类别的集合），相关的算法如下所述。

简单贝叶斯分类算法：

① 计算各个属性的条件概率 $P(C = c_j \mid A_1 = a_1,\ \cdots, A_n = a_n)$

贝叶斯定理：$P(c_j \mid a_1, a_2,\cdots, a_n)$ $= \dfrac{P(a_1, a_2, \cdots, a_n \mid c_j) P(c_j)}{P(a_1, a_2, \cdots, a_n)}$

$$= P(a_1, a_2, \cdots, a_n \mid c_j) P(c_j)$$

属性独立：$P(a_1, a_2,\cdots, a_n \mid c_j)$ $= \prod_{i=1}^{n} P(a_i \mid c_j)$

② 预测推论新测试样本所应归属的类别：

$$c_{NB}=\arg\max_{c_j\in C}P(c_j\mid a_1,a_2,\cdots,a_n)=\arg\max_{c_j\in C}P(c_j)\prod_i P(a_i\mid c_j)$$

综上所述，只要简单贝叶斯分类所涉及的属性相互独立的条件被满足时，简单贝叶斯分类所得到的最大可能分类结果 c_{NB}，和贝叶斯定理的最大可能假说 h_{MAP} 的结果是一致的。

以下例说明简单贝叶斯分类如何进行概念学习，并进行分类预测：

某银行希望能增加办理信用卡的人数。假设目前考虑办卡的相关属性有“性别”、“年龄”、“学生身份”、“收入”四种。分类目标为“办卡”，类别有“会”、“不会”两种，假设现有如表 8-1 所示的 10 笔训练样本。则根据表 8-1 所示，使用简单贝叶斯分类，会将女性，年龄介于 31～45 之间，不具学生身份，收入中等的个人归类到“会”办理信用卡的类别中。

表 8-1　10 笔训练样本

项　目	性　别	年　龄	学生身分	收　入	办　卡
1	男	>45	否	高	会
2	女	31～45	否	高	会
3	女	20～30	是	低	会
4	男	<20	是	低	不会
5	女	20～30	是	中	不会
6	女	20～30	否	中	会
7	女	31～45	否	高	会
8	男	31～45	是	中	不会
9	男	31～45	否	中	会
10	女	<20	是	低	会

要判断（女性，年龄介于 31～45 之间，不具学生身份，收入中等者）会不会办理信用卡，首先应根据训练样本，计算各属性在不同分类结果下的条件概率：

P(性别 ＝ 女|办卡 ＝ 会) = 5/7　　　　P(性别 ＝ 女|办卡 ＝ 不会) = 1/3

P(年龄 ＝31～45|办卡 ＝ 会) = 3/7　　　　P(性别 ＝31～45|办卡 ＝ 不会) = 1/3

P(学生 ＝ 否|办卡 ＝ 会) = 5/7　　　　P(学生 ＝ 否|办卡 ＝ 不会) = 0/3

P(收入 ＝ 中|办卡 ＝ 会) = 2/7　　　　P(收入 ＝ 中|办卡 ＝ 不会) = 2/3

应用简单贝叶斯分类进行类别预测：

$$c_{NB}=\arg\max_{c_j\in\{会,不会\}}P(c_j)\prod_i P(a_i\mid c_j)$$

$$=\arg\max_{c_j\in\{会,不会\}}P(c_j)P(性别=女\mid c_j)P(年龄=31\sim45\mid c_j)$$

$$\times P(学生=否\mid c_j)P(收入=中\mid c_j)$$

再计算有关的条件概率值：

P(办卡 ＝ 会) = 7/10

P(办卡 ＝ 不会) = 3/10

P(会)P(女|会)P((31～45)|会)P(否|会)P(中|会) = 15/343≈0.044

$P(\text{不会})P(\text{女}|\text{不会})P(31\sim45|\text{不会})P(\text{中}|\text{不会}) = 0$

因此基于表 8-1 的训练样本，对于女性，年龄介于 31～45 之间，不具学生身份，收入中等的个人，简单贝叶斯分类会将其分类到会办理信用卡的类别。而且办理的概率是 (0.044)/(0.044 + 0) = 1（正规化分类的结果 $P(\text{会})/(P(\text{会})+P(\text{不会}))$）。

简单贝叶斯分类对于各种属性相对于目标值（分类的类别）的条件概率，是先找出训练样本中，某目标值出现的个数（n），及在这些目标值的样本中，特定属性值出现的个数 n_a，然后将 n_a/n 作为该特定属性在该目标值下的条件概率。如上例 $P(\text{性别} = \text{女}|\text{办卡} = \text{会})$ 的条件概率是 5/7，因为 10 笔训练样本一共有 7 笔是会办卡，而会办卡的 7 笔中，有 5 笔是女性。

因为各属性间是相互独立的，一旦有一个条件概率为零，这种方法计算出来的各项目标值都是零。上例不会办卡的概率为零，因为受了 $P(\text{学生} = \text{否}|\text{办卡} = \text{不会}) = 0$ 的影响，不会办卡的概率就为零了。为了克服训练样本选取不够广泛造成零概率的困境，简单贝叶斯分类采用了 m-estimate 加以改良，从而能更精确地作出适当的分类。m-estimate 的定义为：

$$M = \frac{n_a + mp}{n + m}$$

其中：

m 是一个固定的常数值，主要用来决定 p 的权重；

p 为同一属性不同属性值的事前概率，一般而言采用均匀分布的概率值，如上例性别只有两种可能，均值的概率，使得 p=1/2。

8.2 Excel 2007 贝叶斯概率分类

Step1：选择【数据挖掘】→【高级】→【创建挖掘模型】命令，如图 8-1 所示。

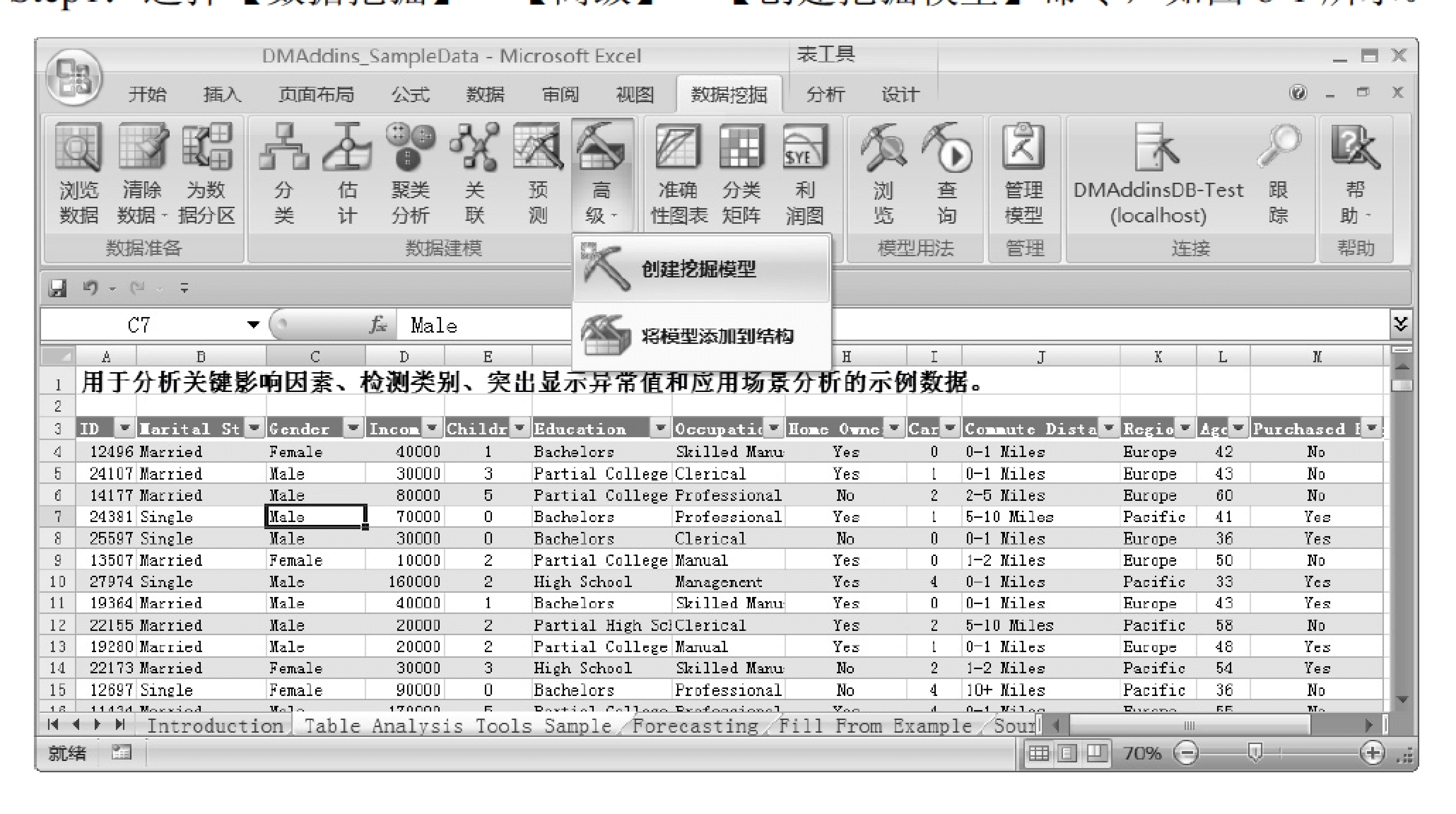


图 8-1 创建挖掘模型

Step2：开始使用创建模型向导，单击【下一步】按钮，如图 8-2 所示。

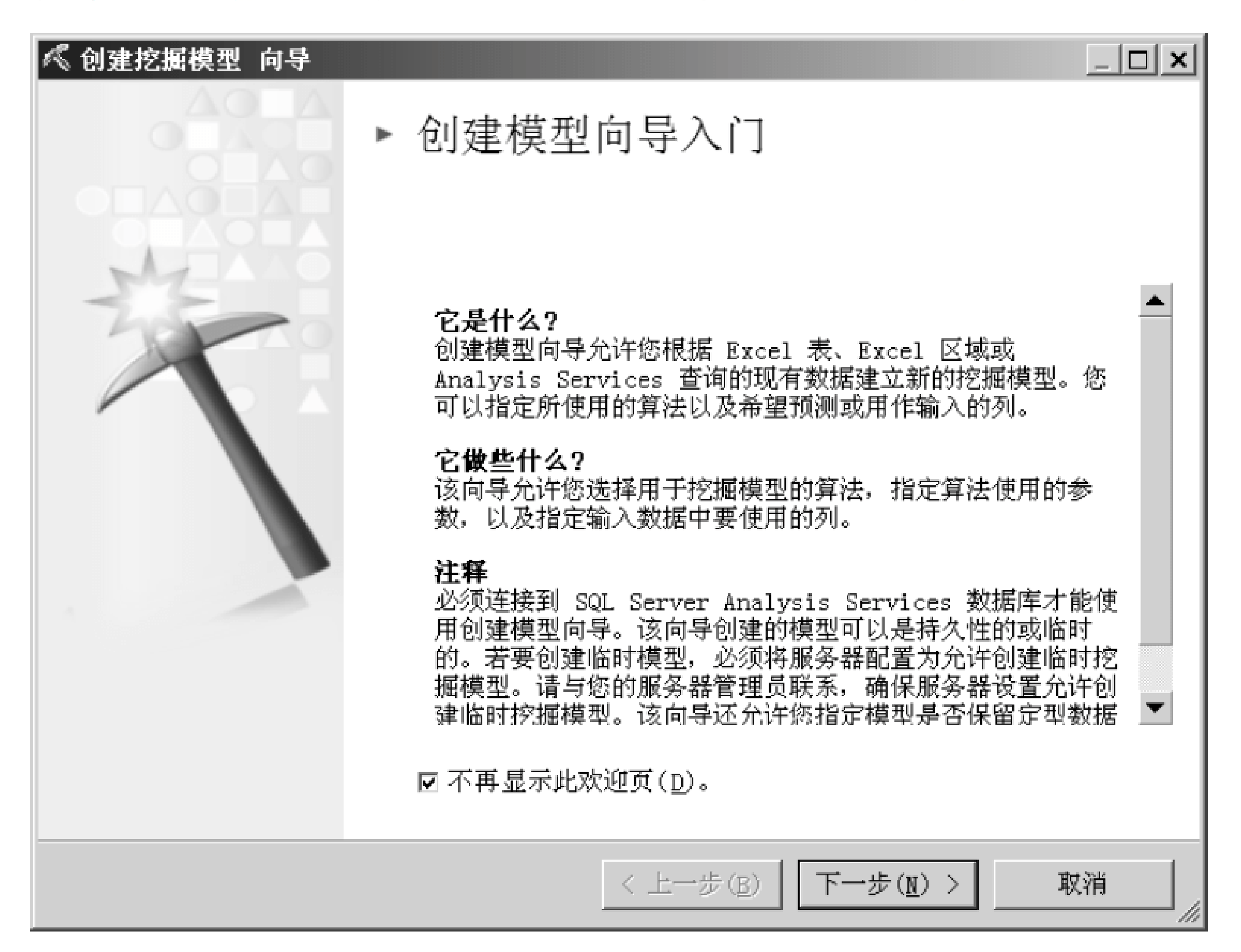


图 8-2　创建模型向导

Step3：选中【数据区域】单选按钮，选中【我的数据区域包含页眉】复选框，并选择红色部分可选择数据区域，如图 8-3 所示。



图 8-3　选择数据区域

Step4：在【区域选择】文本框中选择或者输入特定的数据范围，如图 8-4 所示。



图 8-4　输入特定的数据范围

Step5：单击【下一步】按钮，如图 8-5 所示。



图 8-5　选择源数据

Step6：在【算法】下拉列表框中选择 Microsoft Naive Bayes 选项，如图 8-6 所示。单击【参数】按钮可更改其参数预设值，这里采用其默认设定，不做修改。



图 8-6　选择挖掘算法

Step7：将自变量设定为“输入”，预测变量设定为“仅预测”，如图 8-7 所示。此数据中的序号设定为 key，若不使用的变量则设定为“不使用”，完成后单击【下一步】按钮。



图 8-7　选择列

Step8：选中【浏览模型】复选框，单击【完成】按钮，可更改【结构名称】及【模型名称】文本框内容，如图 8-8 所示。

Step9：选择【依赖关系网络】选项卡，若结果有多个变量与预测变量存在关系，则可调整所有链接，找出其中关联的强弱程度，如图 8-9 所示。



图 8-8　完成

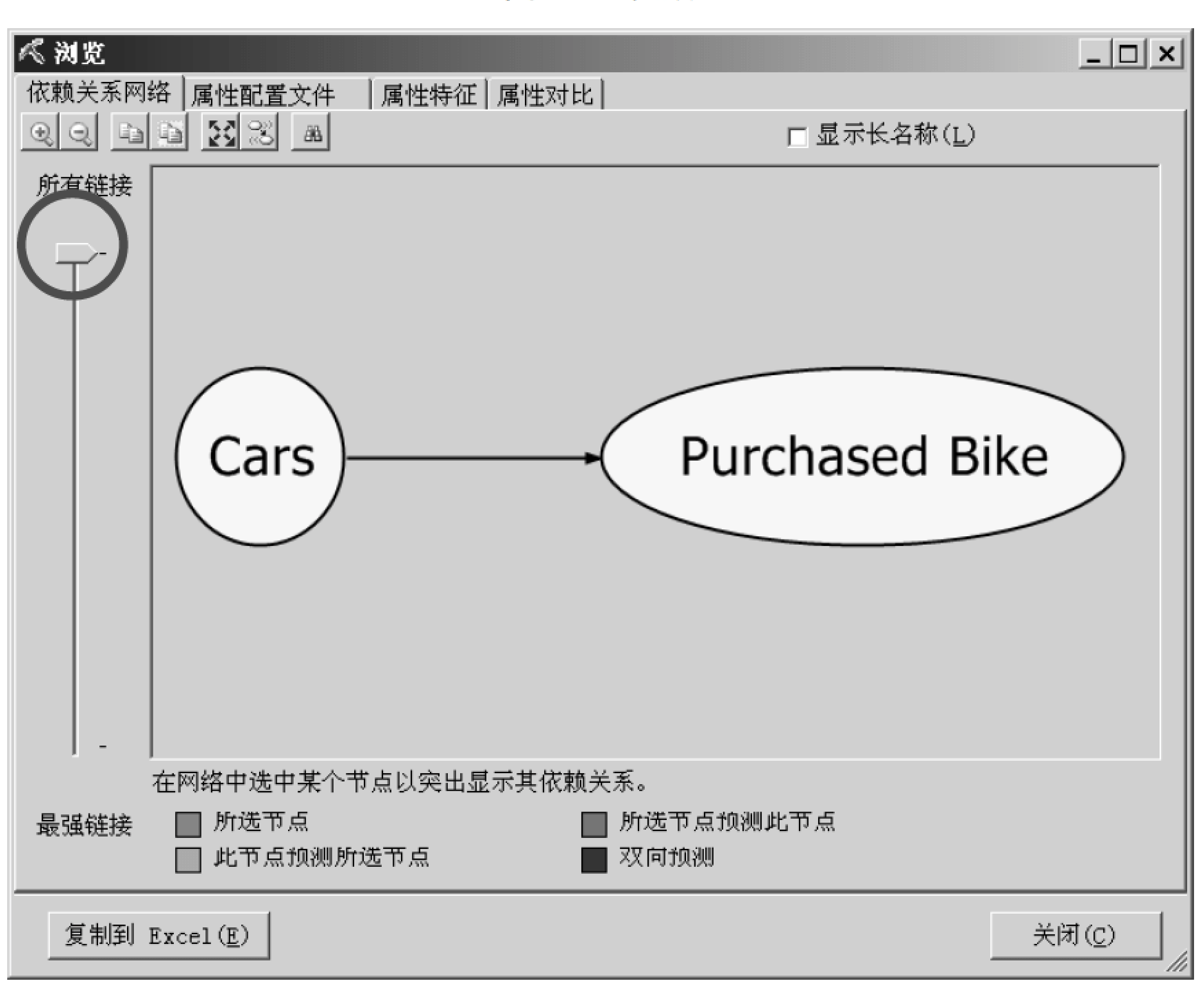


图 8-9　【依赖关系网络】选项卡

Step10：选择【属性配置文件】选项卡，可调整【直方图列】，如图 8-10 所示。

Step11：单击【复制到 Excel】按钮可将结果输出到 Excel 中，从如图 8-11 所示中可得知所选出的自变量在不同状态下与预测变量的结果比较。



图 8-10 【属性配置文件】选项卡

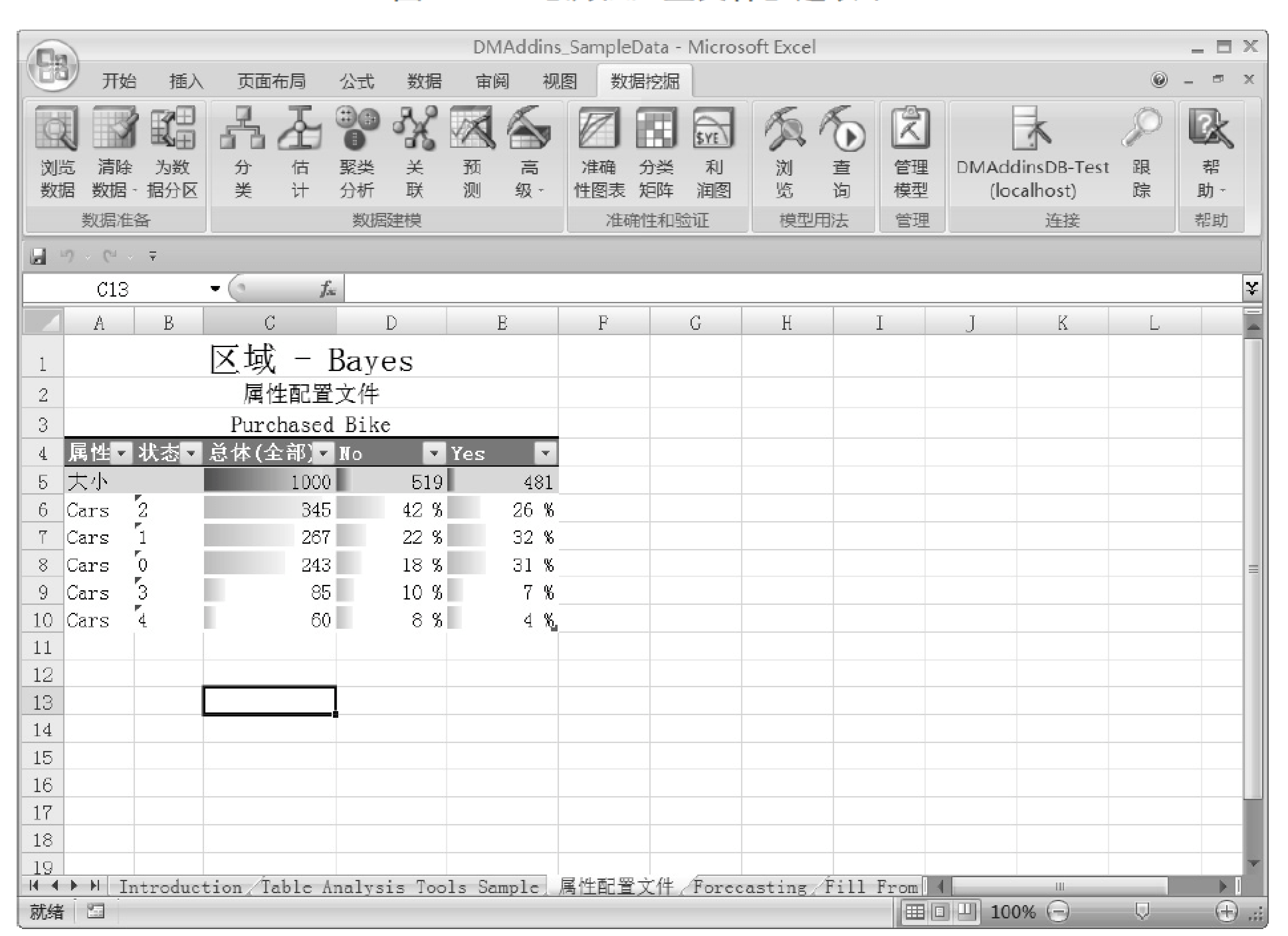


图 8-11 将结果输出到 Excel

Step12：选择【属性特性】选项卡，可更改【值】为不同的预测变量，单击【复制到 Excel】按钮，如图 8-12 所示。

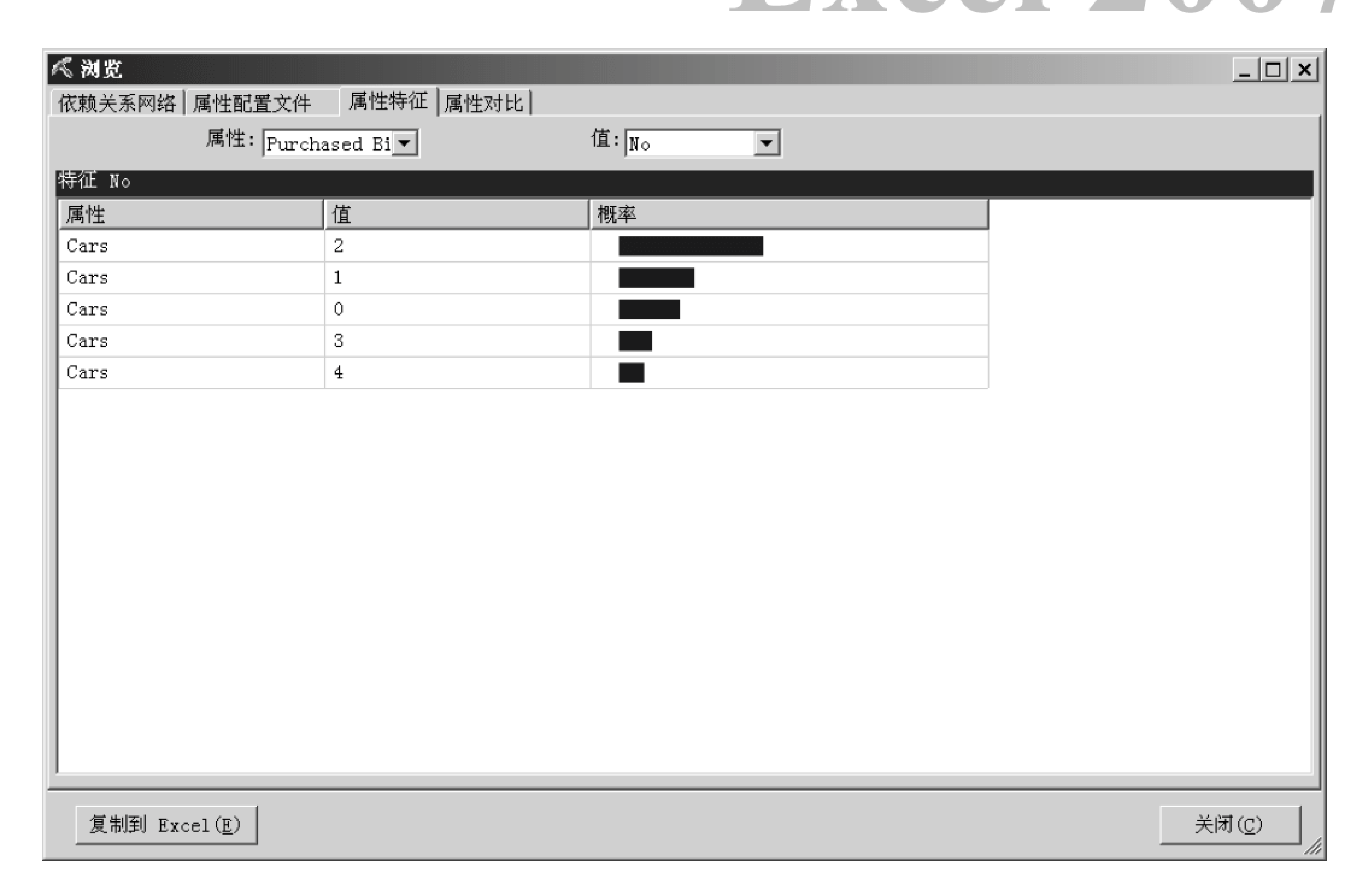


图 8-12 【属性特征】选项卡

Step13：可将 Step12 中所列出的结果显示在 Excel 上，如图 8-13 所示。

区域 - Bayes

属性特征

Purchased Bike=Yes

属性	值	概率
Cars	1	32 %
Cars	0	31 %
Cars	2	26 %
Cars	3	7 %
Cars	4	4 %

图 8-13 在 Excel 上显示的结果

Step14：选择【属性对比】选项卡，调整【值 1】及【值 2】下拉列表框内容，如图 8-14 所示。



图 8-14 【属性对比】选项卡

Step15：单击【复制到 Excel】按钮，可将 Step14 中所列结果显示在 Excel 上，如图 8-15 所示。

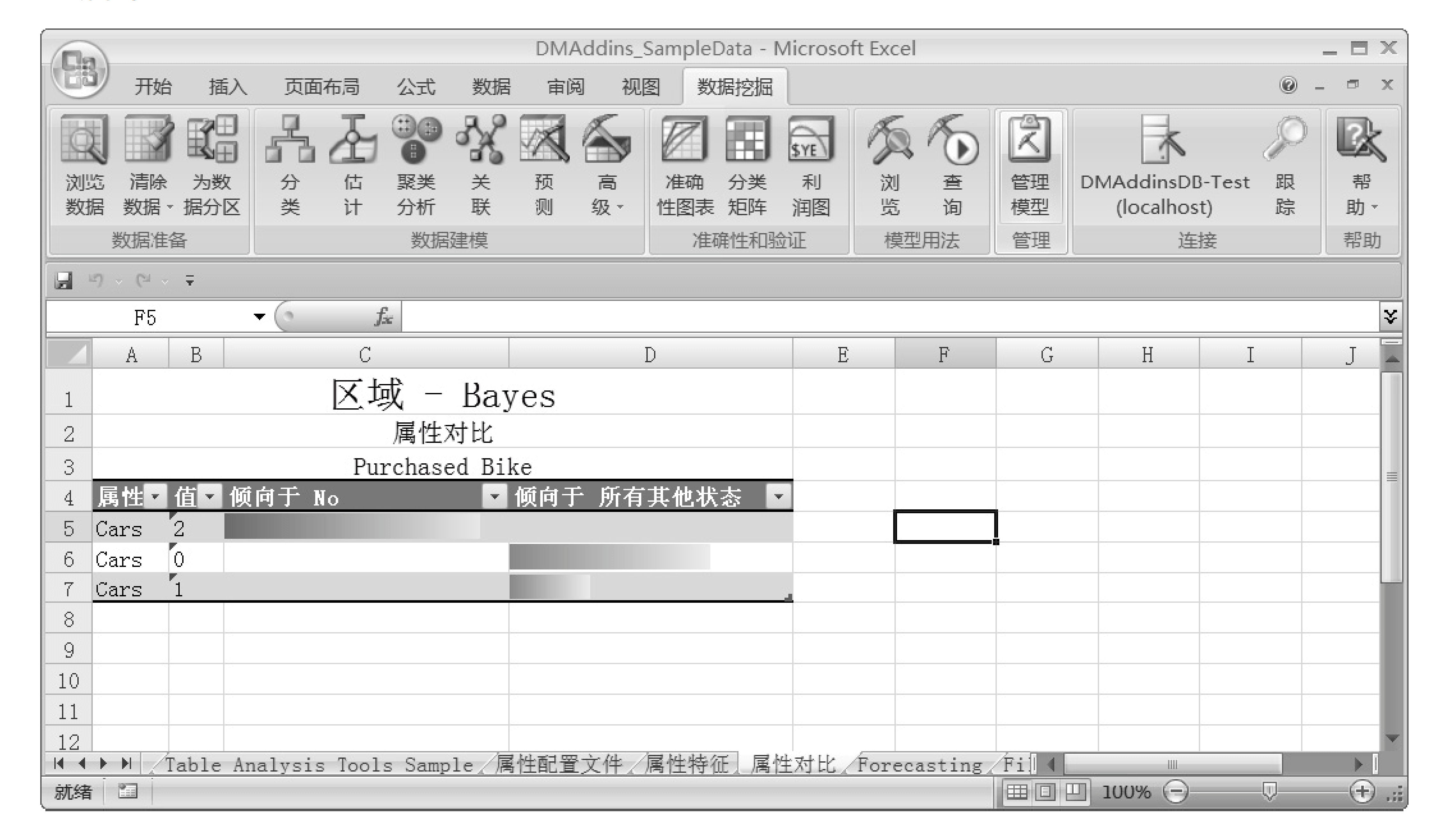


图 8-15 在 Excel 上显示的结果

Step16：单击【数据挖掘】中的【利润图】按钮，弹出【利润图向导入门】窗口，然后单击【下一步】按钮，如图 8-16 所示。

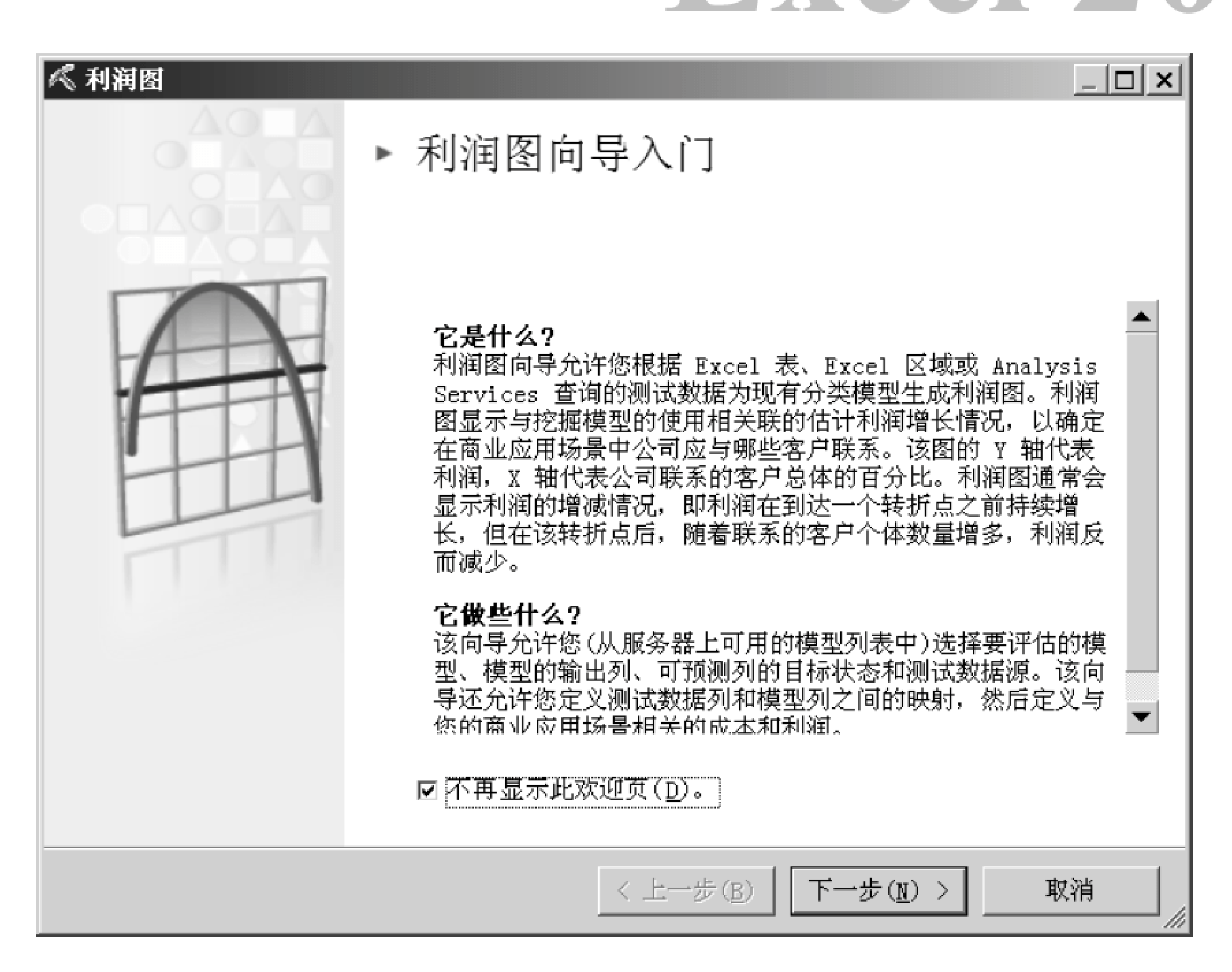


图 8-16 【利润图向导入门】窗口

Step17：选择【区域结构】中的贝叶斯，单击【下一步】按钮，如图 8-17 所示。

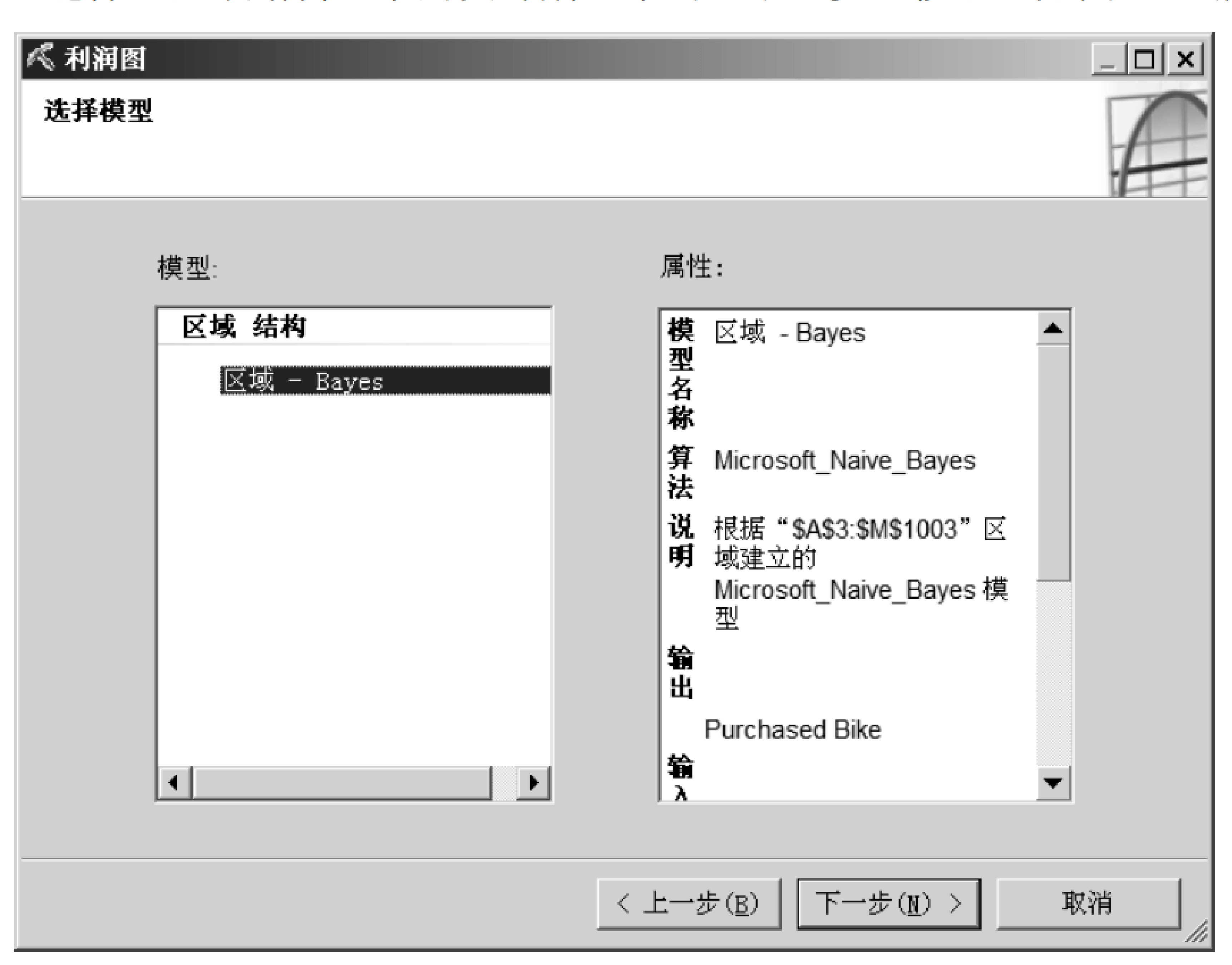


图 8-17 选择模型

Step18：可调整“要预测的值”、“目标总体”、“固定成本”、“单项成本”、“单项收入”等项目。调整完后单击【下一步】按钮，如图 8-18 所示。

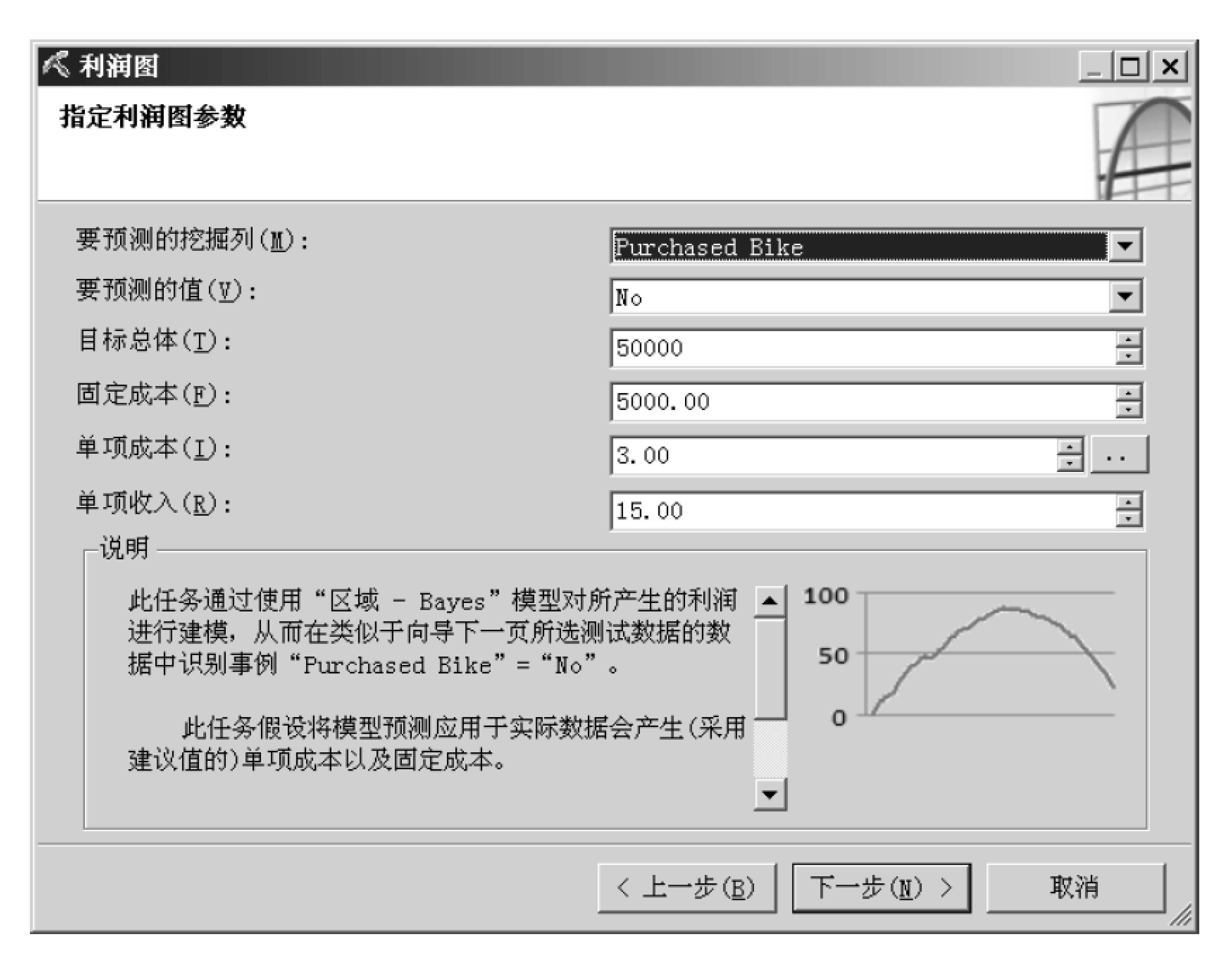


图 8-18　调整参数

Step19：选中【数据区域】单选按钮，并选中【我的数据区域包含页眉】复选框，选择红色部分可选择数据，如图 8-19 所示。



图 8-19　选择源数据

Step20：单击【下一步】按钮，弹出如图 8-20 所示的【指定关系】窗口。



图 8-20 【指定关系】窗口

Step21：单击【完成】按钮后，可得到该模型的利润图，如图 8-21 所示。

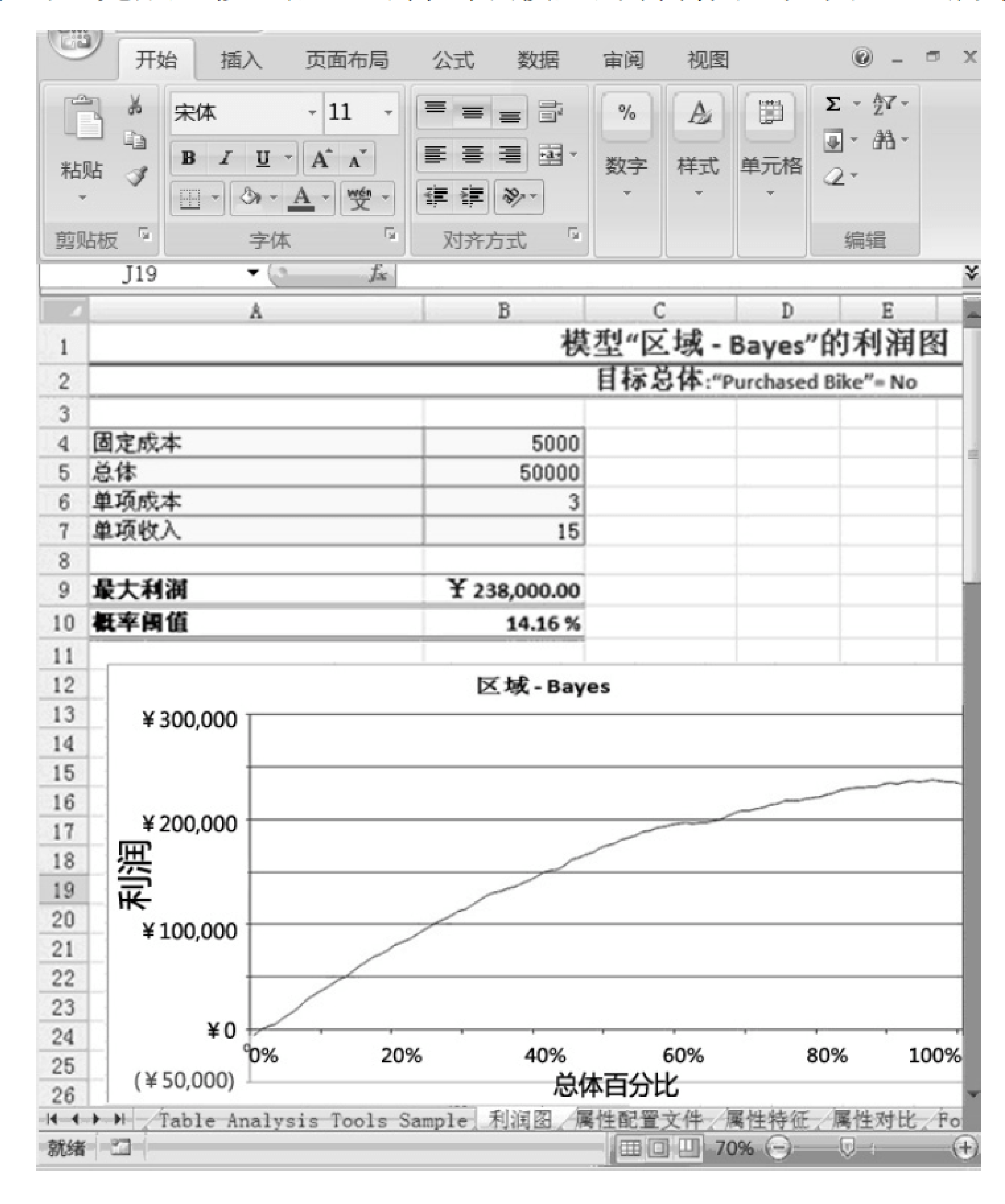


图 8-21 利润图

第 9 章　关 联 规 则

9.1　基 本 概 念

关联规则是分析并发现数据库中不同变量或个体间（例如研究不同商品间的关系及年龄与购买行为等）的关联程度（或概率大小），并利用关联规则建立顾客购买行为模型，如购买了台式计算机对购买其他计算机外设商品（打印机、音箱、硬盘等）的相关影响。发现这些规则可以应用于商品货架摆设、库存安排以及根据购买行为模型对客户进行分类等。

关联规则最早是由 Agrawal 于 1993 年提出，他对关联规则的定义如下：

假设 $I=\{I_1,I_2,\cdots,I_n\}$：I 可视为 m 个商品项目的集合。

$D=\{t_1,t_2,\cdots,t_n\}$：D 为 n 位客户交易的总集合。

其中 $t_i=\{I_{i1},I_{i2},\cdots,I_{ik}\}$：$t_i$ 代表第 i 位客户的交易数据。

关联规则的代表式“if *condition* then *result* ”。即：“$X \Rightarrow Y$”，其中 X、Y 称作项目集（itemsets）。

关联规则中有两个重要的参数，分别为支持度（support）和可信度（confidence）。其中支持度是指 X 项目集与 Y 项目集，同时出现在 D 交易总集合的次数，除以 D 交易总集合的个数；以概率的观点来看，支持度就是同时发生 X、Y 事件的概率。可信度是指 X 项目集与 Y 项目集，同时出现在 D 交易总集合的次数，除以 X 项目集在 D 交易总集合出现的次数；以概率的观点来看，可信度就是在 X 事件发生的情况下，Y 事件发生的概率。

例如：有商品牛奶和面包，其被购买的概率如表 9-1 所示。

表 9-1　商品被购买概率

事 件 组 合	概率/%
牛奶	35
面包	50
牛奶和面包	25

得到的关联规则为：“牛奶⇒面包”支持度为 0.25，可信度为 0.25/0.35=0.714。意思是全部顾客中，有 25%的人买了牛奶也买了面包，而且买牛奶这项商品的顾客中，有 71.4%的人也会一起购买面包。

另外，有些学者认为单以支持度和可信度衡量规则的好坏不够充分，还需考虑项目集彼此间的相互关系。因此又产生了“兴趣度”（interesting）或称“增益”（improvement）等指标，其具体的公式如下：

$$兴趣度=\frac{\text{Confident}(X\Rightarrow Y)}{P(Y)}=\frac{P(X\,\&\,Y)}{P(X)P(Y)}$$

当兴趣度大于 1 时，这条规则就是比较好的；当兴趣度小于 1 时，这条规则就是没有太大意义的。兴趣度越大，规则的实际意义就越好。

9.2 关联规则的种类

将关联规则按不同的情况进行分类。

1．基于规则中处理的变量的类别，关联规则可以分为布尔型和数值型

布尔型关联规则处理的值都是离散的、类别的，它显示了这些变量之间的关系；而数值型关联规则可以和多维关联或多层关联规则结合起来，对数值型字段进行处理，将其进行动态地分割，或者直接对原始的数据进行处理，当然数值型关联规则中也可以包含种类变量。

例如：性别=“女”⇒职业=“秘书”，是布尔型关联规则；性别=“女”⇒avg（收入）=2 300，涉及的收入是数值类型，所以是一个数值型关联规则。

2．基于规则中数据的抽象层次，可以分为单层关联规则和多层关联规则

在单层的关联规则中，所有的变量都没有考虑到现实的数据是具有多个不同的层次的；而在多层的关联规则中，对数据的多层性已经进行了充分考虑。

例如：IBM 台式机⇒Sony 打印机，是一个细节数据上的单层关联规则；台式机⇒Sony 打印机，是一个较高层次和细节层次之间的多层关联规则。

3．基于规则中涉及的数据的维数，关联规则可以分为单维的和多维的

在单维的关联规则中，只涉及数据的一个维，如用户购买的物品；而在多维的关联规则中，要处理的数据将会涉及多个维。换句话说，单维关联规则是处理单个属性中的一些关系；多维关联规则是处理各个属性之间的某些关系。

例如：啤酒⇒尿布，这条规则只涉及用户购买的物品；性别=“女”⇒职业=“秘书”，这条规则就涉及两个字段的信息，是两个维上的一条关联规则。

给出了关联规则的分类之后，在下面的分析过程中，就可以考虑某个具体的方法适用于哪一类规则的挖掘，某类规则又可以用哪些不同的方法进行处理。

9.3 关联规则的算法：Apriori 算法

Apriori 算法为研究关联规则的入门算法，也是研究关联规则最具代表性的算法之一。其利用迭代的方式，找出数据库中项目集的并发关系，并形成规则。

1．执行步骤

① 首先，须指定最小支持度及最小可信度。

② Apriori 算法使用了候选项目集的观念，首先产生出项目集，称为候选项目集，若候选项目集的支持度大于或等于最小支持度，则该候选项目集为高频项目集(large itemset)。

③ 在 Apriori 算法的过程中，首先由数据库读入所有的交易，得出候选单项目集（candidate 1-itemset）的支持度，再找出高频单项目集（large 1-itemset），并利用这些高频单项目集的结合，产生候选 2 项目集（candidate 2-itemset）。

④ 再扫描数据库，得出候选 2 项目集的支持度以后，再找出高频 2 项目集，并利用这些高频 2 项目集的结合，产生候选 3 项目集。

⑤ 重复扫描数据库，与最小支持度比较，产生高频项目集，再结合产生下一级候选项目集，直到不再结合产生出新的候选项目集为止。

2．优点

简单易懂，容易实现。

3．缺点

因计算项的个数过多而造成执行缓慢，主要原因在于高频项目集产生过多的候选项目集，尤其是候选 2 项目集的情况最为严重，因为相当于计算所有的项目集。

9.4 Excel 2007 关联规则

Step1：使用 Excel 2007 SQL 2005 DM addin 范例，数据为范例数据中的 Associate 窗口的数据。单击工具栏中的【关联】按钮，弹出如图 9-1 所示的【关联向导入门】窗口，单击【下一步】按钮。

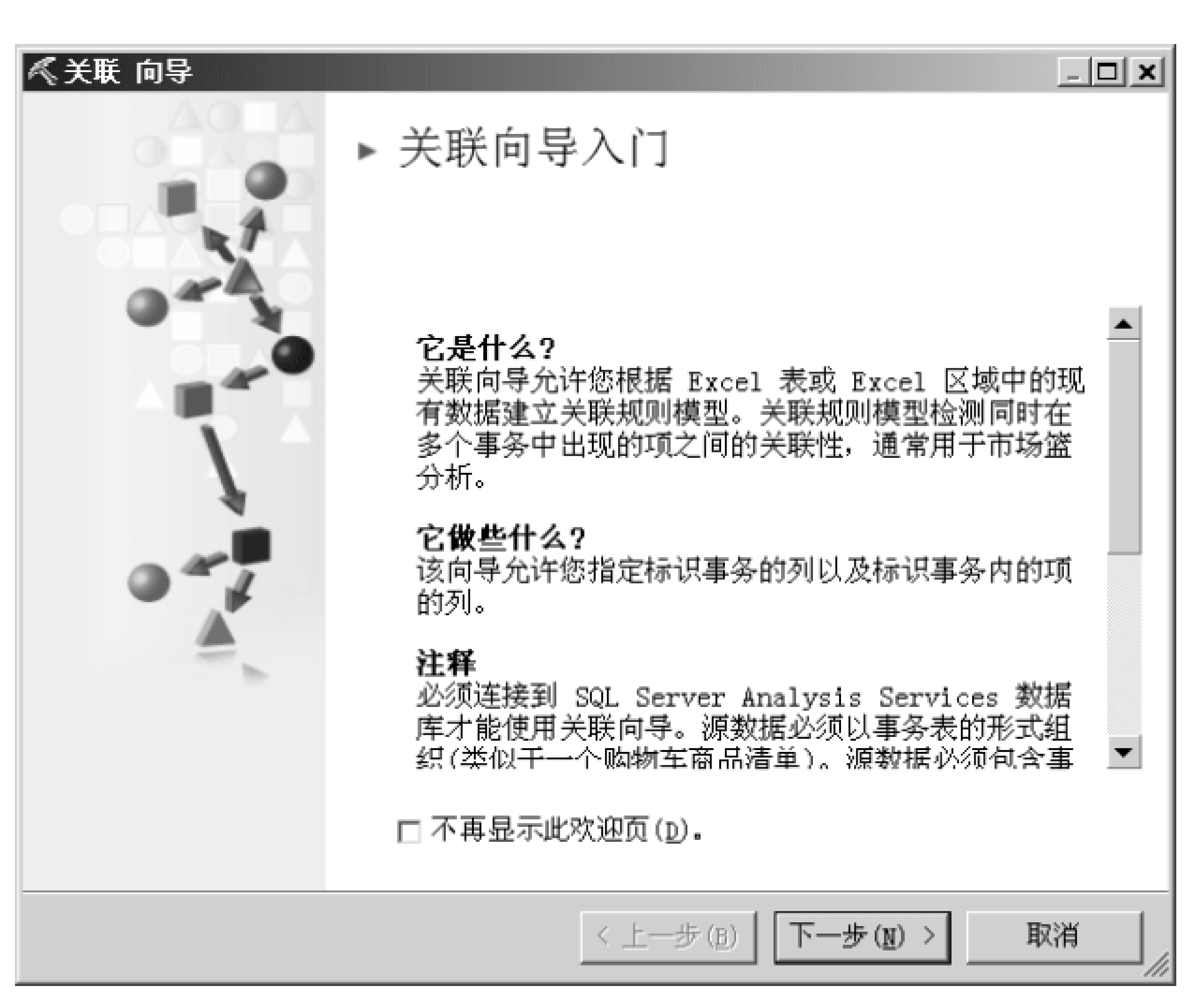


图 9-1 【关联向导入门】窗口

Step2：选择表或者选择数据区域，单击【下一步】按钮，如图 9-2 所示。



图 9-2　选择源数据

Step3：决定所要预测的项目，在此预测类别目录与产品之间的关联，然后单击【下一步】按钮，如图 9-3 所示。

图 9-3　关联

Step4：选中【启用钻取】复选框，并单击【完成】按钮，如图 9-4 所示。



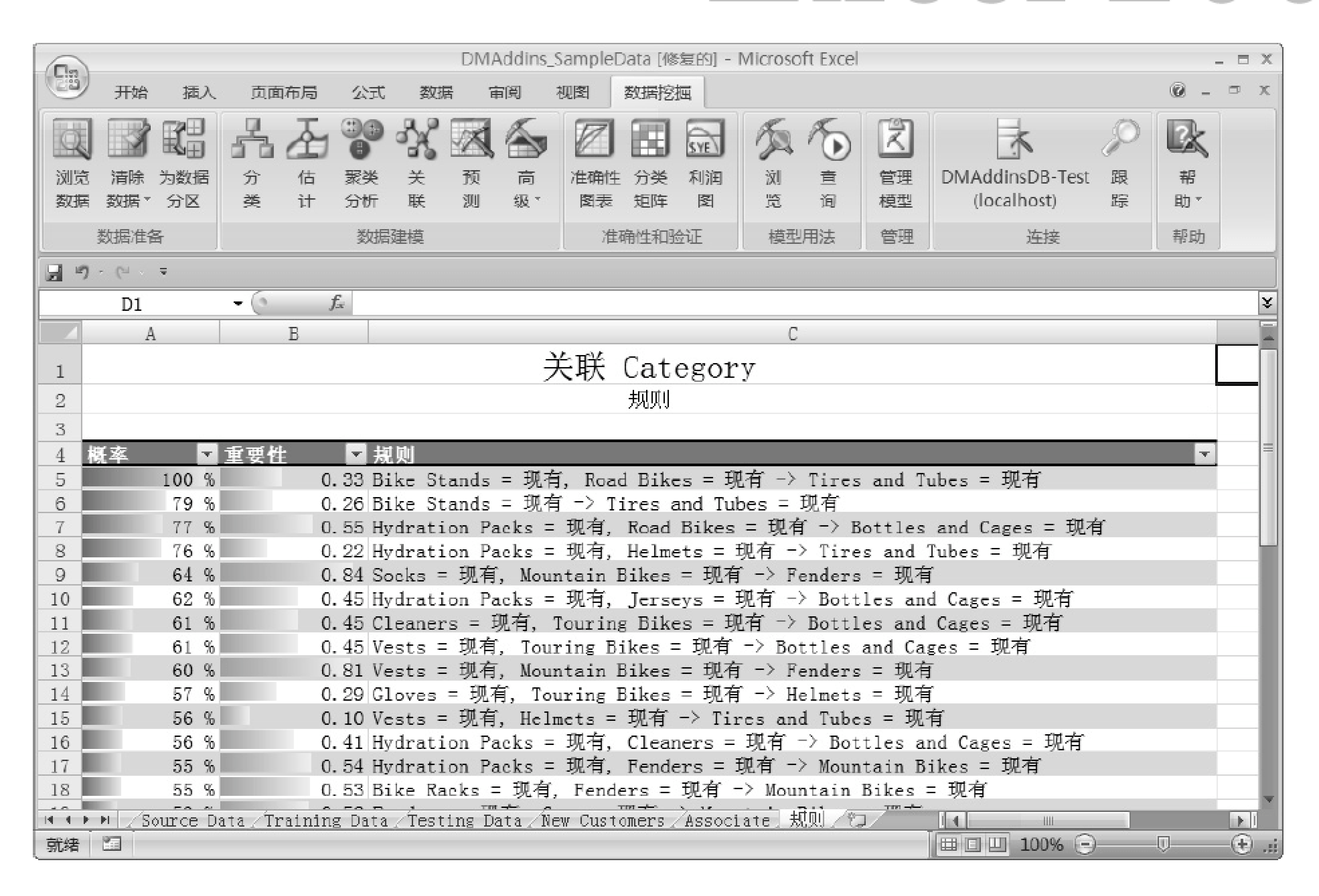
图 9-4　单击【完成】按钮

Step5：关联规则如图 9-5 所示，图中展示了各种类别目录或产品间的关联，也同时给出了其概率值和重要性。



图 9-5　关联规则

Step6：将图表复制到 Excel，如图 9-6 所示。



概率	重要性	规则
100 %	0.33	Bike Stands = 现有, Road Bikes = 现有 -> Tires and Tubes = 现有
79 %	0.26	Bike Stands = 现有 -> Tires and Tubes = 现有
77 %	0.55	Hydration Packs = 现有, Road Bikes = 现有 -> Bottles and Cages = 现有
76 %	0.22	Hydration Packs = 现有, Helmets = 现有 -> Tires and Tubes = 现有
64 %	0.84	Socks = 现有, Mountain Bikes = 现有 -> Fenders = 现有
62 %	0.45	Hydration Packs = 现有, Jerseys = 现有 -> Bottles and Cages = 现有
61 %	0.45	Cleaners = 现有, Touring Bikes = 现有 -> Bottles and Cages = 现有
61 %	0.45	Vests = 现有, Touring Bikes = 现有 -> Bottles and Cages = 现有
60 %	0.81	Vests = 现有, Mountain Bikes = 现有 -> Fenders = 现有
57 %	0.29	Gloves = 现有, Touring Bikes = 现有 -> Helmets = 现有
56 %	0.10	Vests = 现有, Helmets = 现有 -> Tires and Tubes = 现有
56 %	0.41	Hydration Packs = 现有, Cleaners = 现有 -> Bottles and Cages = 现有
55 %	0.54	Hydration Packs = 现有, Fenders = 现有 -> Mountain Bikes = 现有
55 %	0.53	Bike Racks = 现有, Fenders = 现有 -> Mountain Bikes = 现有

图 9-6　复制到 Excel

Step7：关联项目集，如图 9-7 所示。



支持	大小	项集
12	3	Bike Racks = 现有, Fenders = 现有, Mountain Bikes = 现有
13	3	Bike Stands = 现有, Road Bikes = 现有, Tires and Tubes = 现有
27	3	Bottles and Cages = 现有, Helmets = 现有, Tires and Tubes = 现有
67	3	Caps = 现有, Bottles and Cages = 现有, Helmets = 现有
15	3	Caps = 现有, Bottles and Cages = 现有, Tires and Tubes = 现有
127	3	Caps = 现有, Helmets = 现有, Tires and Tubes = 现有
78	3	Caps = 现有, Jerseys = 现有, Bottles and Cages = 现有
73	3	Caps = 现有, Jerseys = 现有, Helmets = 现有
51	3	Caps = 现有, Jerseys = 现有, Mountain Bikes = 现有
79	3	Caps = 现有, Jerseys = 现有, Road Bikes = 现有
66	3	Caps = 现有, Jerseys = 现有, Tires and Tubes = 现有
74	3	Caps = 现有, Mountain Bikes = 现有, Bottles and Cages = 现有
49	3	Caps = 现有, Mountain Bikes = 现有, Helmets = 现有
18	3	Caps = 现有, Mountain Bikes = 现有, Tires and Tubes = 现有
70	3	Caps = 现有, Road Bikes = 现有, Bottles and Cages = 现有
65	3	Caps = 现有, Road Bikes = 现有, Helmets = 现有
24	3	Caps = 现有, Road Bikes = 现有, Tires and Tubes = 现有
50	3	Caps = 现有, Touring Bikes = 现有, Bottles and Cages = 现有
43	3	Caps = 现有, Touring Bikes = 现有, Helmets = 现有
31	3	Caps = 现有, Touring Bikes = 现有, Jerseys = 现有
14	3	Caps = 现有, Touring Bikes = 现有, Tires and Tubes = 现有
11	3	Cleaners = 现有, Bottles and Cages = 现有, Tires and Tubes = 现有
21	3	Cleaners = 现有, Fenders = 现有, Mountain Bikes = 现有
13	3	Cleaners = 现有, Fenders = 现有, Tires and Tubes = 现有
28	3	Cleaners = 现有, Mountain Bikes = 现有, Bottles and Cages = 现有
28	3	Cleaners = 现有, Mountain Bikes = 现有, Tires and Tubes = 现有

图 9-7　关联项目集

Step8：将图表复制到 Excel，如图 9-8 所示。

Step9：图 9-9 为各关联的依赖关系网络图，由图中可发现大致分为四个部分。

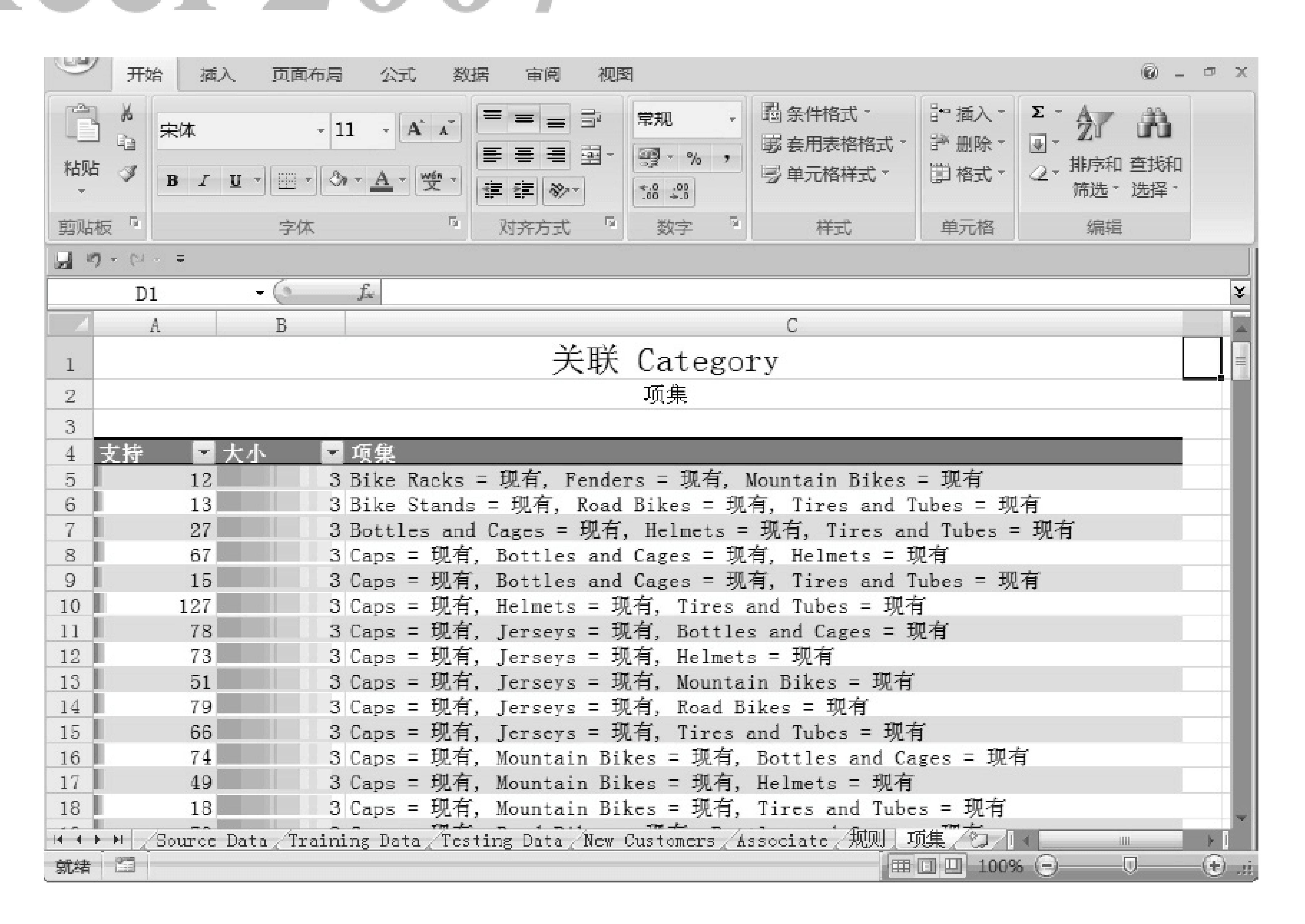


图 9-8　复制到 Excel

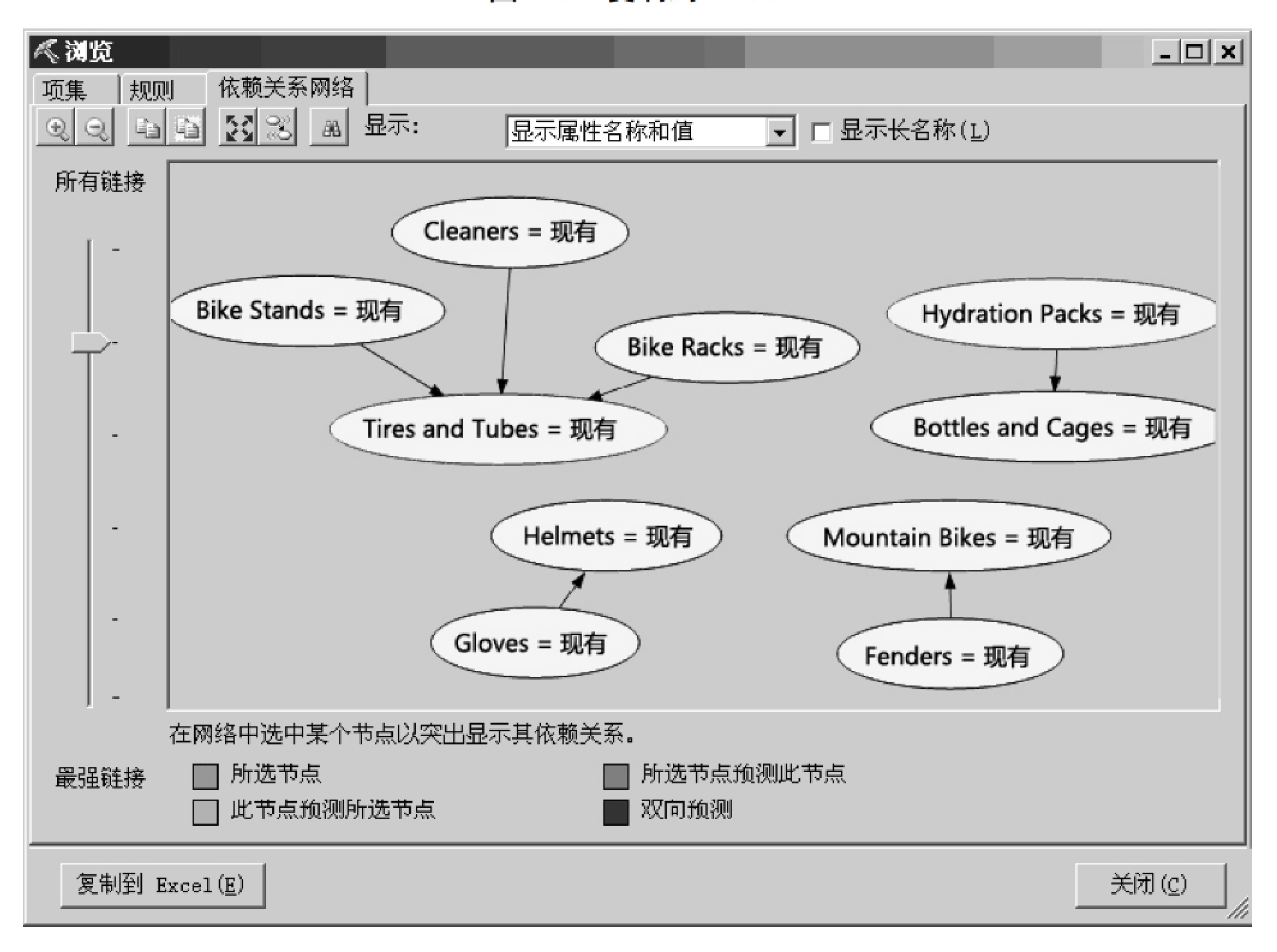


图 9-9　依赖关系网络

Step10：选择关系链接强度，可知道各类别目录或产品的关系强度，其中分别是 Bike Stands 和 Tires and Tubes、Bottles and Cages 和 Hydration Packs、Mountain Bikes 和 Fenders 的关系是最强的，如图 9-10 所示。

Step11：也可选择【高级】→【创建挖掘模型】命令，在此选用范例 Table Analysis Tools Sample 工作表，如图 9-11 所示。



图 9-10　各类别关系强度



图 9-11　创建挖掘模型

Step12：在选择预测关联变量时，选择 Marital Status 为预测变量，如图 9-12 所示。



图 9-12　选择列

Step13：单击【下一步】按钮，并选中【启用钻取】复选框，如图 9-13 所示。



图 9-13　选中【启用钻取】复选框

Step14：图 9-14 为关联项目集。

Step15：将图表复制到 Excel，如图 9-15 所示。

Step16：图 9-16 为关联规则的概率值和重要性。

Step17：将图表复制到 Excel，如图 9-17 所示。



图 9-14　关联项目集

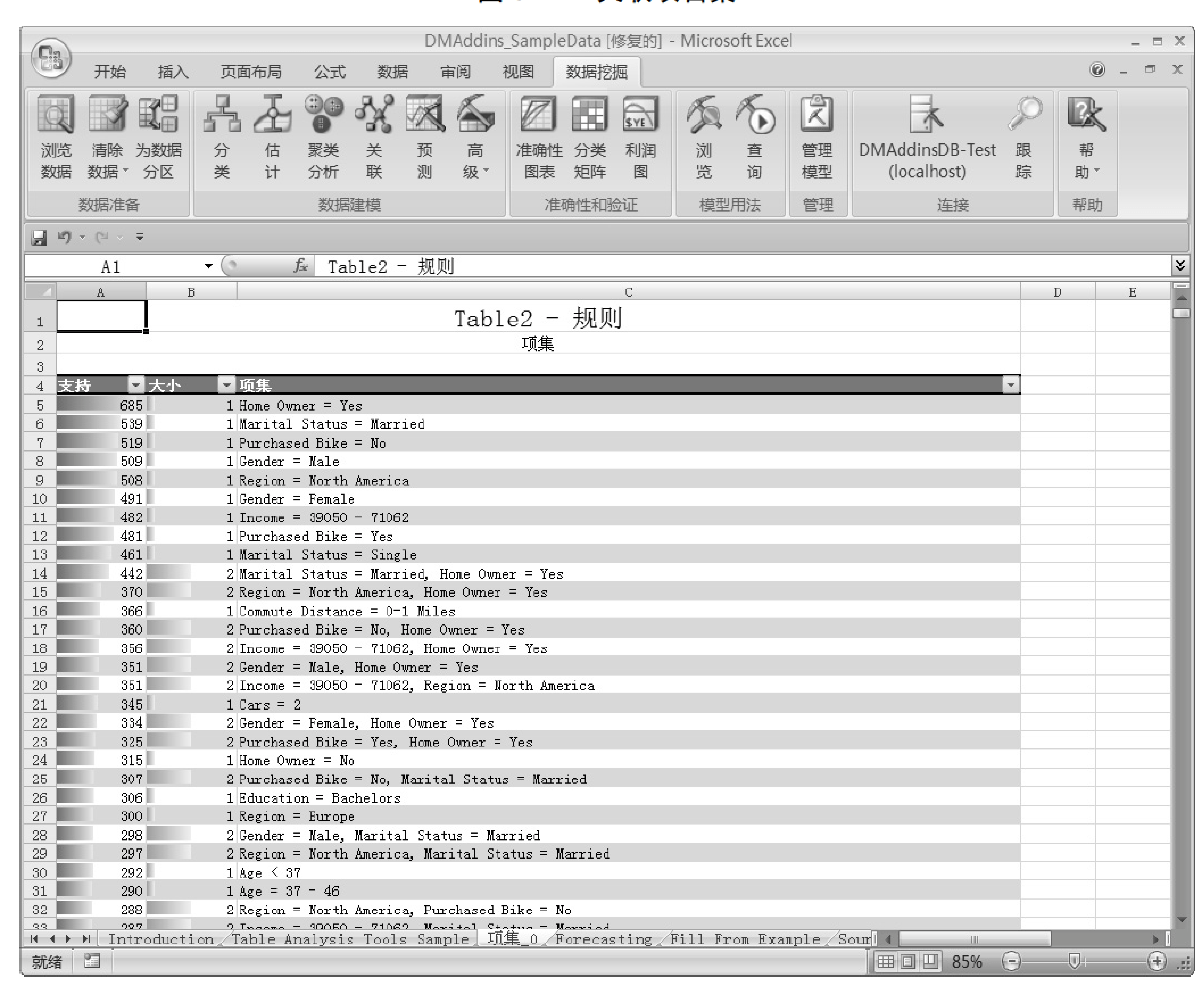


图 9-15　复制到 Excel



图 9-16 关联规则

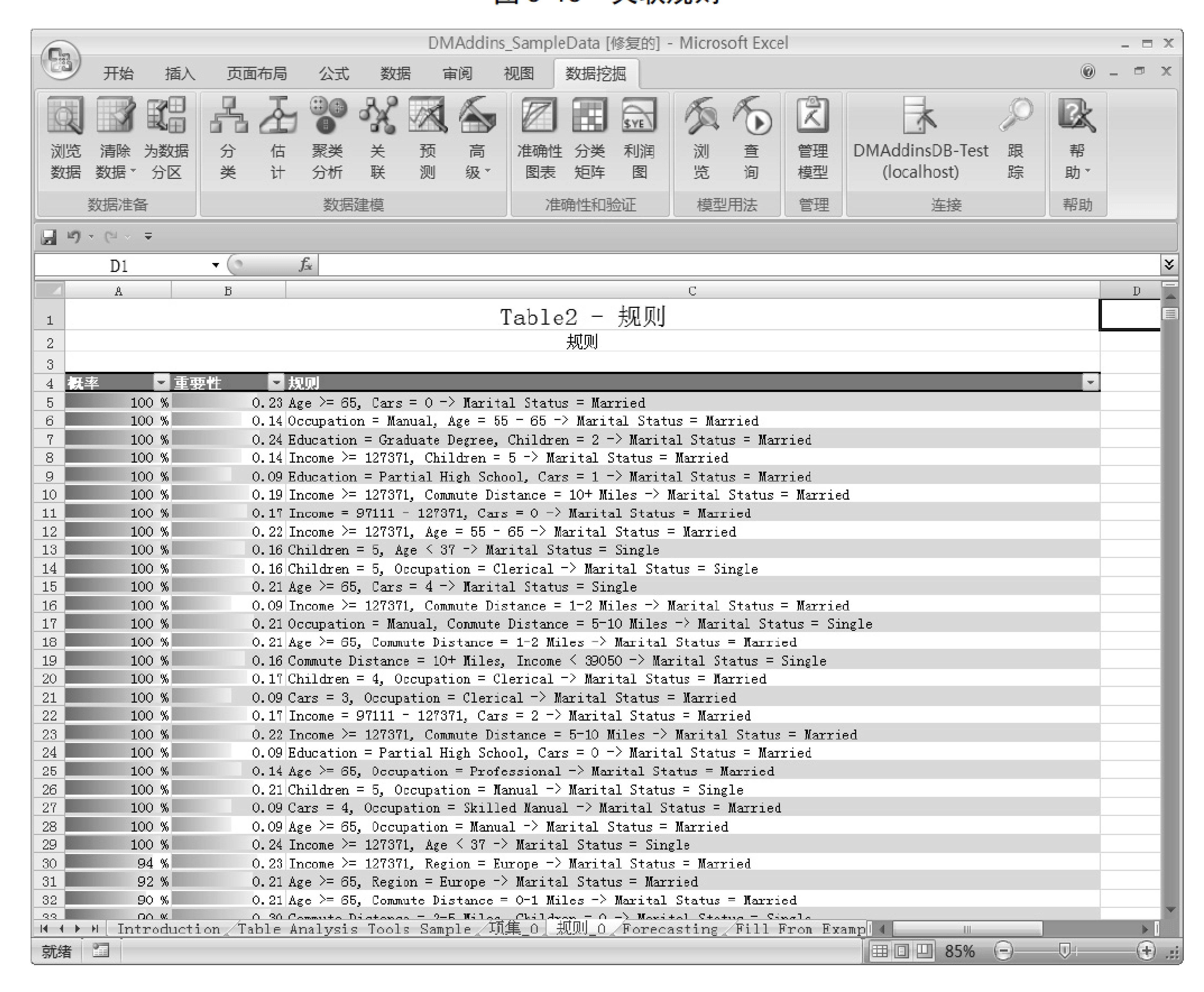


图 9-17 复制到 Excel

Step18：图 9-18 为关联规则的依赖关系网络图。

Step19：将图表复制到 Excel，如图 9-19 所示。

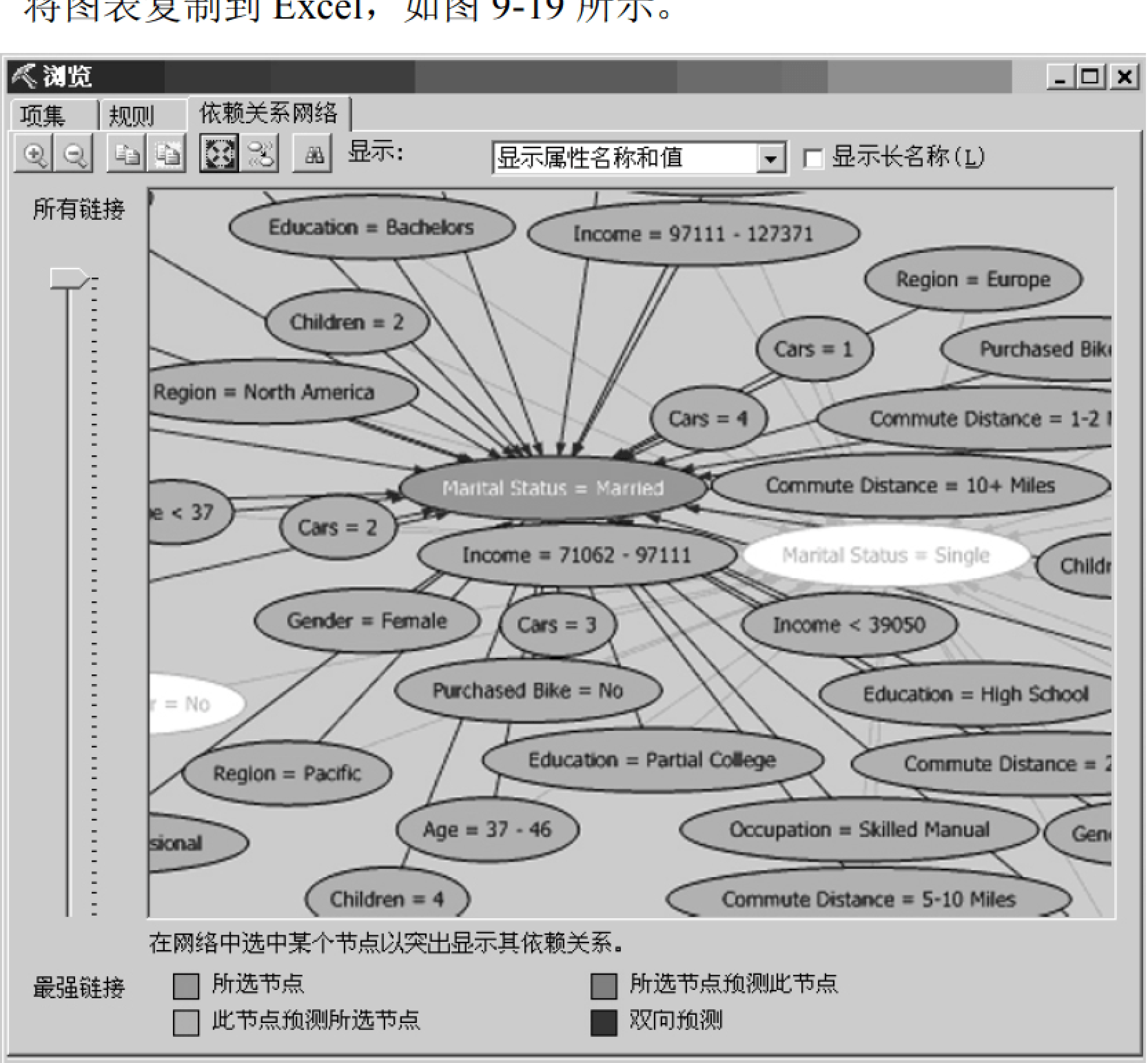



图 9-18 依赖关系网络

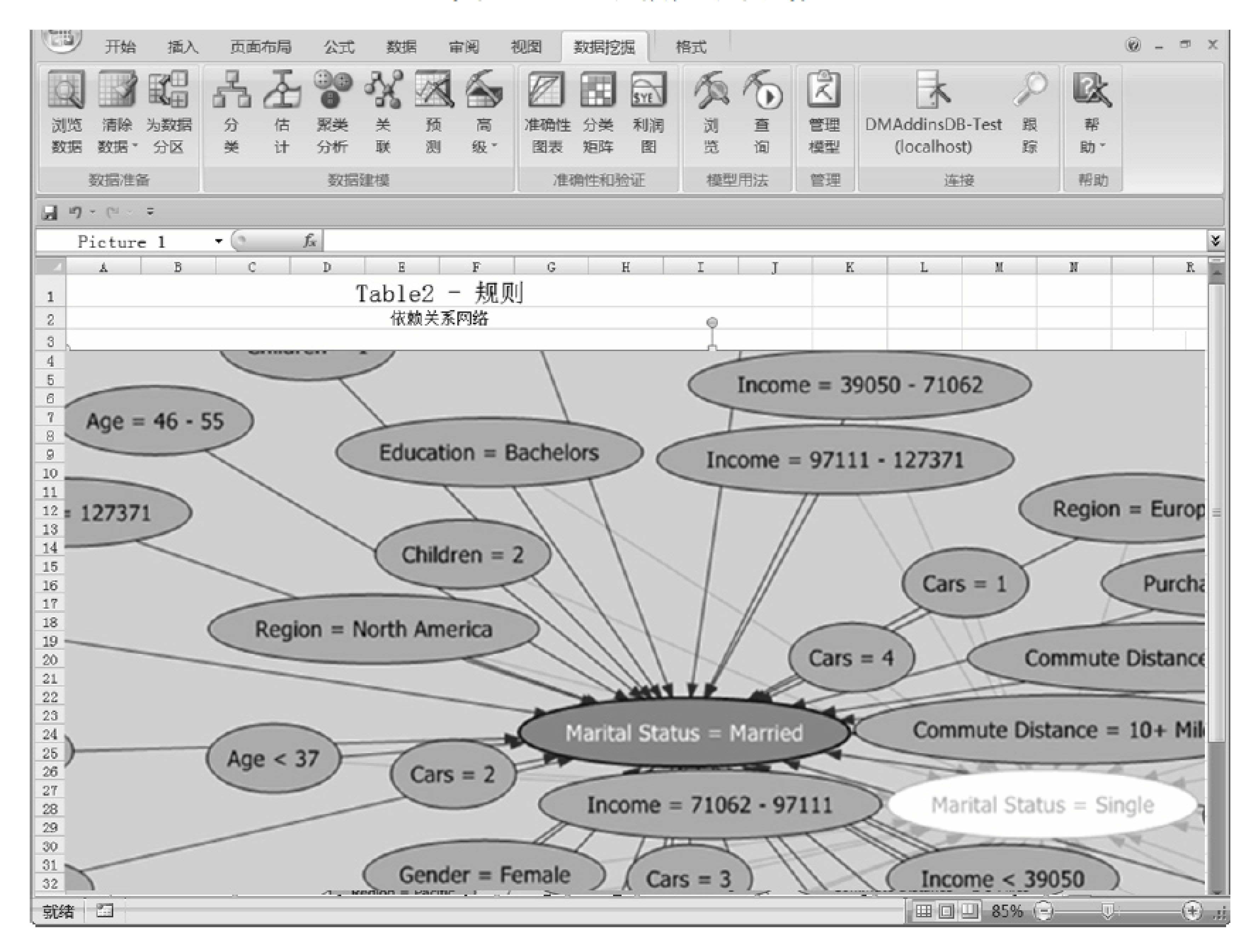


图 9-19 复制到 Excel

第 10 章　聚 类 分 析

10.1　基 本 概 念

聚类分析（cluster analysis）的观念与判别分析非常相似，同样也是希望能够通过对样本分类，寻找到多变量个体的差异。但却有两点不同：①聚类分析的分群方式并不需要预先指定最终的类别个数，完全由算法和数据决定；②聚类分析属于一种非参数分析方法，所以并没有非常严谨的数理依据，当然也不需假设总体为正态分布。聚类分析常用于对数据进行约简或分类，也就是把相似的个体归于一类。不过，究竟相似的标准是什么？多么相似才能归为一类？聚类分析的结果是否有意义？这些都是需要探讨的问题。

没有一种最好的聚类分析方法可以解决所有的问题，因此，在聚类分析时，必须设定分析目标，并根据聚类分析的目标选择各种变量。为了避免变量选择的偏误，也可以使用其他方法加以辅助，如图形法等。通常，聚类分析可以分为以下两个基本步骤。

① 搜集数据（data collection）。在搜集数据时，应先确立分析目标，而后选择有代表性的数据（因为变量空间的形态会影响类别的形态，故必须小心选择），注意选用合适的测量单位。必要的时候需要进行数据变换，例如对数变换、平方根变换、标准化变换、异常值剔除等。

② 转换成相似矩阵（transformation to similary matrix）。由于聚类分析是把相似性大的个体归为一群，所以必须逐个计算出个体间两两相似系数（similation coefficient），并形成相似矩阵（similary matrix）。然后按照相似性程度归并个体为一群。

10.2　层次聚类分析

根据相似性统计量，将样本或变量进行聚类的主要方法有以下几种。

1．系统聚类法

系统聚类法是目前使用最多的一种聚类方法，这种聚类方法是先将聚类的样本或变量各自看成一群，然后确定群与群之间的相似统计量，并选择最接近的两群或若干个群合并成一个新群，计算新群与其他各群之间的相似性统计量，再选择最接近的两群或若干群合并成一个新群，直到所有的样本或变量都合并成一群为止。

常用的系统聚类法是以距离为相似统计量时，确定新群与其他各群间距离的方法，如最短距离法、最长距离法、中间距离法、重心法、群平均法、离差平方法和欧氏距离法等。

2．逐步聚类法

系统聚类法的优点是聚类比较准确，缺点是聚类的次数较多，每聚类一次只能减少一群或若干个群，每一次都需要计算两两样本或各群之间的距离或其他相似性统计量，比较麻烦。

而逐步聚类法相对简便，先确定若干个样本为初始凝聚点，计算各样本与凝聚点的距离或其他相似性统计量；进行初始聚类后，再根据初始聚类计算各群的重心作为新的凝聚点，进行第二次聚类；再给出一个初始的聚类方案，再按照某种最优法则，逐步调整聚类方案，直到得到最优的聚类方案。用逐步聚类法解题的关键是凝聚点的选择及聚类结果的调整，常用的方法有成批调整法、逐个调整法及离差平方和法。

3．逐步分解法

这种方法是先将所有的个体看成一群，然后反复对现有的群进行分解，直到各个群都不能分解为止。

4．有序样本的聚类

这种方法适用于有顺序的对象，聚类后既保持了个体原有的顺序，又按照某种最优法则分割为若干个互有差异的类别。

10.3　聚类分析原理

聚类分析中的相似性，是依据样本点在几何空间上的距离来判断的。样本点之间距离越近，其相似程度就越高，于是就可以归并成为同一组。为了说明的方便起见，以入学申请的 Toefl 与 Gmat 成绩为例。当这些数据转换成几何空间的图像时，可以得到如图 10-1 所示的结果。

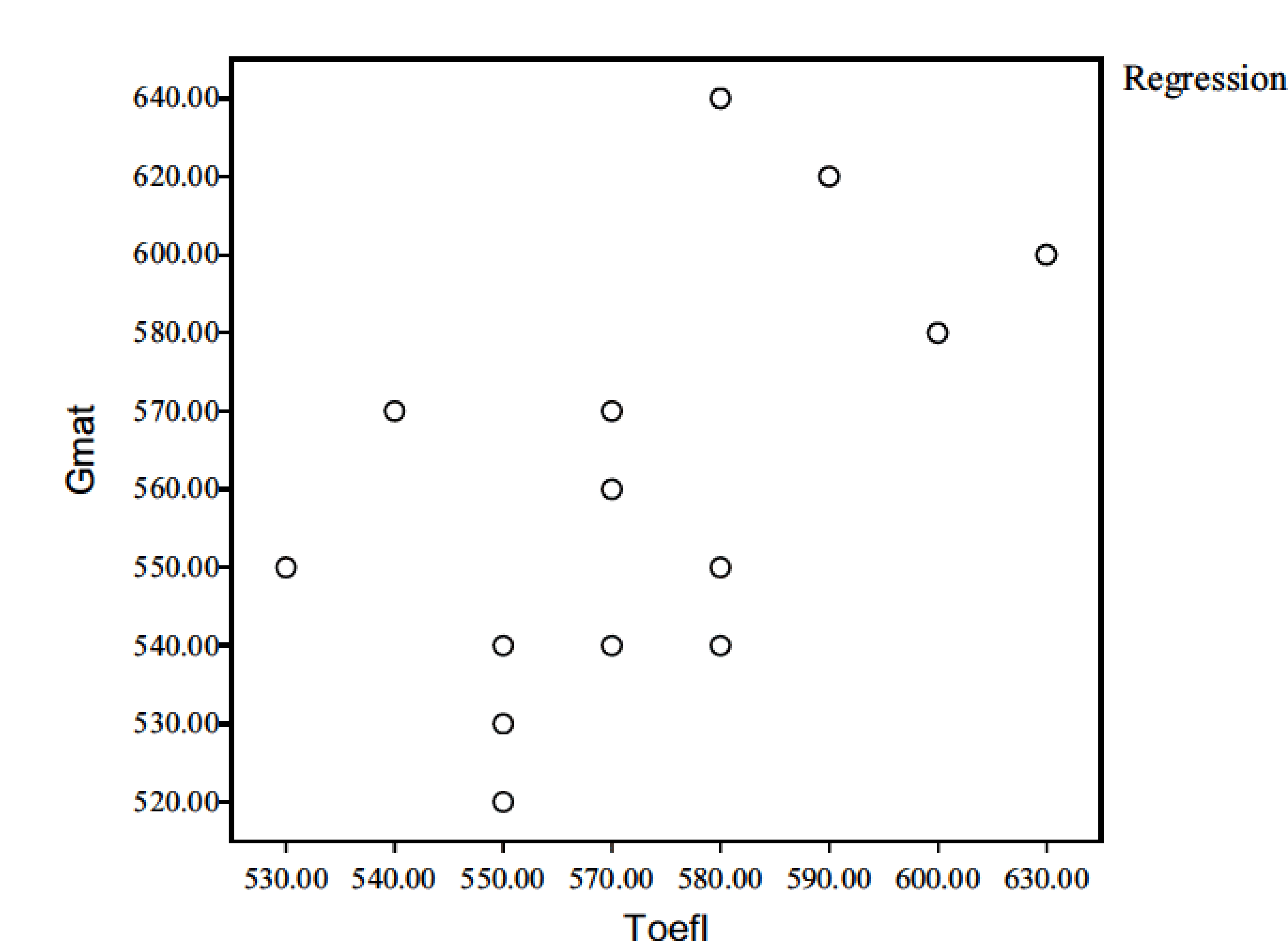
从这个图当中，可以大概地主观归类，把学生划分成左下角与右上角的两个区块。于是将#14、#6、#4、#5 归为一类，其余的学生归为另一类。像这样的划分方法，其实就是利用距离的观念，将距离比较偏远的#14、#6、#4、#5，从多数聚类的聚类当中区分开来（注：此为 cluster seed 观念）。当然也可以反其道而行，就是使用归并的方法，首先将#3 与#11 这两组分数完全相同的学生合并成一组，然后再考虑如何去合并出下一个聚类。

在数学上对于距离这个观念，可以有下列几种不同的定义：

- euclidean 距离：$d_{ij}=\left[\left(x_i-x_j\right)'\left(x_i-x_j\right)\right]^{1/2}=\left[\sum(x_i-x_j)^2\right]^{1/2}$。
- mahalanobis 距离：$D_{ij}^2=[x_i-x_j]'S^{-1}[x_i-x_j]$。
- City block 距离：$d_{ij}=\left|x_i-x_j\right|'\cdot 1=\sum\left|x_i-x_j\right|$。

一般的计算机软件大多使用欧氏距离，作为聚类分析“距离”的计算基础。欧氏距离所衡量出来的是确实的实际距离，例如对于申请人#1 与#2 而言，其欧氏距离的计算方式为：

$$d_{12}=\left[(580-530)^2+(550-550)^2\right]^{1/2}=50$$



ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Toefl	580	530	570	600	630	590	570	580	570	540	570	550	550	580	550
Gmat	550	550	570	580	600	620	540	540	560	570	570	520	530	640	540

图 10-1 入学申请 Toefl 和 Gmat 成绩分布图

欧氏距离适合使用在单位一致、不必加权的多变量数据上。例如使用同一测量尺度的抽样问卷数据。不过对于具有不同单位的数据，例如经济数据当中的人口与所得，具有六位数字以上的数据，与利率通货膨胀率等仅具有小数点以下的数据相结合，其欧氏距离将会被离群值所影响。

马氏距离类似于欧氏距离，但须经过协方差阵的修正，即一般统计观念当中标准化的程序。由于马氏距离也同时考虑到协方差的大小，所以对于距离的衡量，与未经过标准化的欧氏距离作比较时，当然会有差异。正因为如此，利用马氏距离或欧氏距离，来做聚类分析的结果就应该有所不同。也就是说经过标准化的马氏距离，在变量之间相关系数为零时，才有可能与经过标准化后的欧氏距离衡量结果一致。就整体而言，以上马氏或欧氏衡量的差异，在多变量的各个数据非常相近，而协方差阵的差异又颇大时尤其明显。

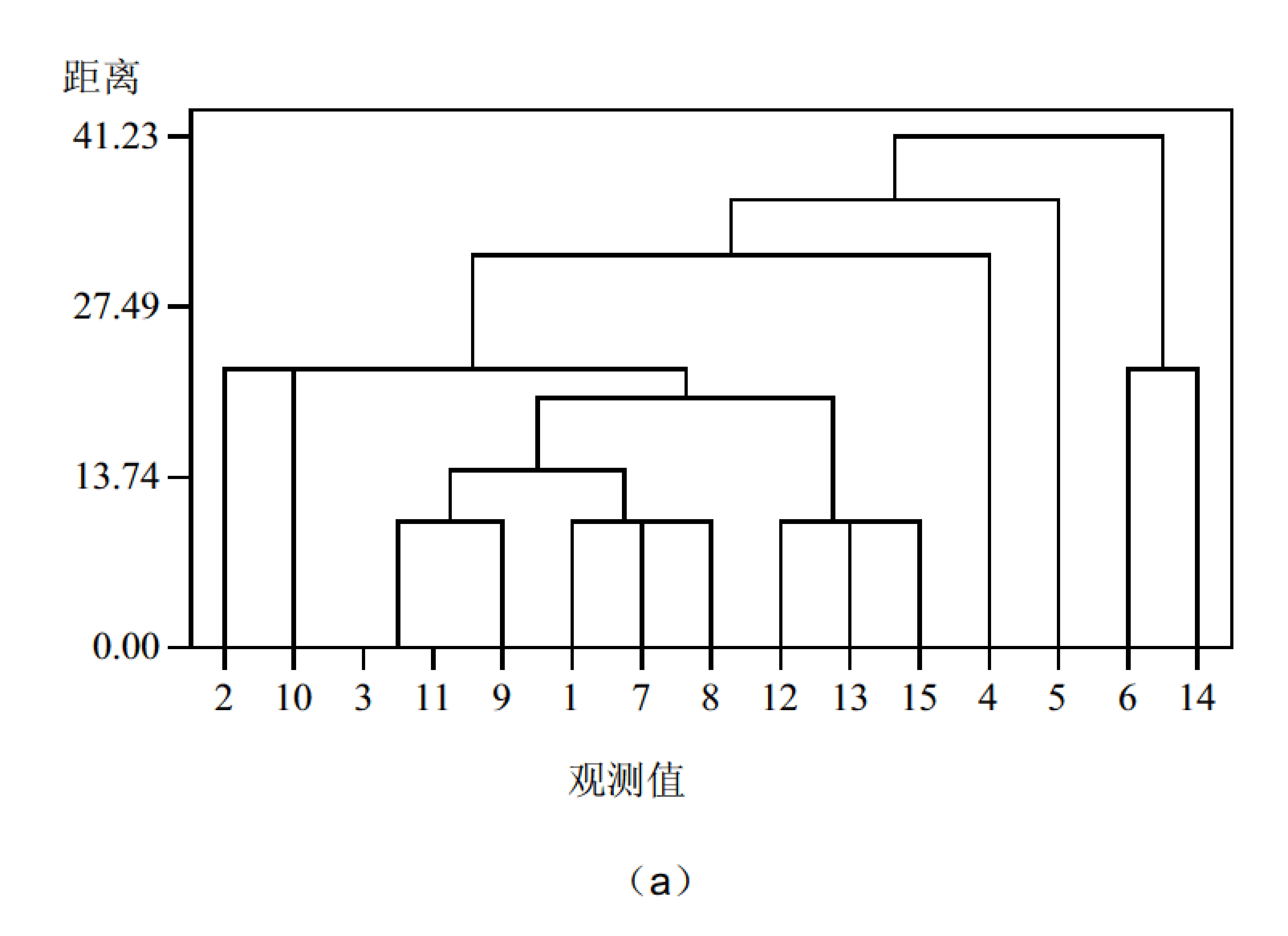
绝对值距离（也称为街区距离）以数据差异的绝对值作为衡量的依据。由于对数据差异没有经过开方与平方根的调整，也不需经过协方差阵的修正，所以依据绝对值距离作聚类分析的结果，当然与前两者会产生相当的差异。它的优点，尤其是对于拥有许多小数点以下变量的数据群特别有用。试想，一个 0.05 的数据差异，经过欧氏或马氏距离的计算之后，平方后的数据是 0.002 5，其分子项会变小。所以不论是欧氏距离还是马氏距离，都有低估比例数据的倾向。当然马氏距离还具有方差作调整的功能，尚不至于产生偏误。

如果仅使用 Toefl 与 Gmat 的分数计算欧氏距离，以作为衡量学生聚类分析的依据时，可以得到如图 10-2 所示的结果。

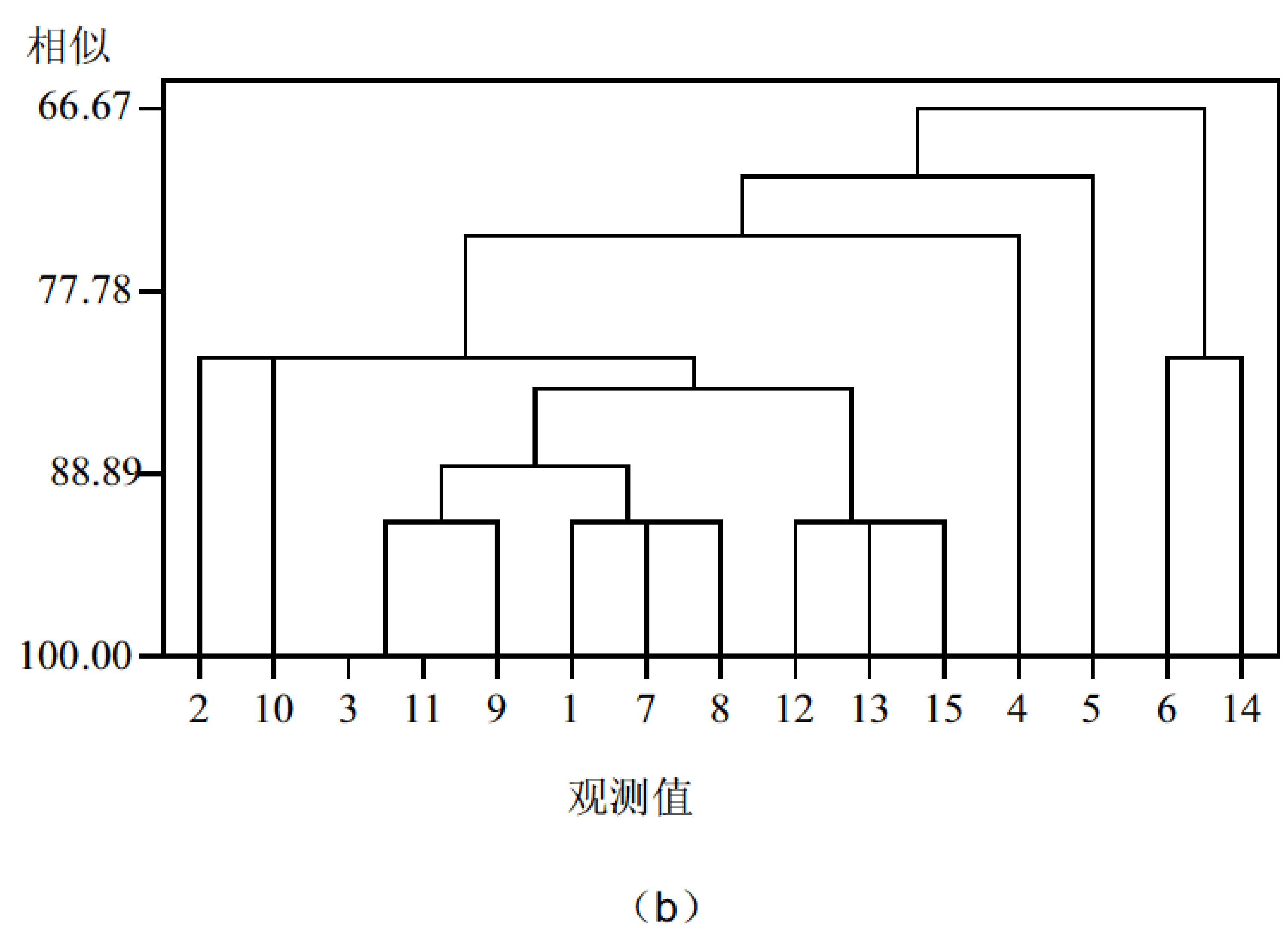
图 10-2（a）所展示聚类的树型图（dendrogram），由下而上展示各个相似的，或者说距离相近的个体，两两相归并的过程。

每次个体的合并都需要付出代价，即图 10-2（b）纵坐标所展示的组内个体距离的增加，

或者相似性的降低。



（a）



（b）

图 10-2 使用 Toefl 与 Gmat 对于申请人的聚类分析结果

这个归并过程，先是由每一个体为中心，再逐步合并最近距离的样本。此方法在聚类分析当中称为层次聚类法（也叫系统聚类法）当中的凝聚法（agglomerative method）。在两两归并的过程当中，聚类的中心点会因为不同的样本值而不断作改变，并且在图像当中不断地移动位置。

若希望中心点不要因为两两合并的过程而改变，必须使用不同于层次聚类法的非层次聚类法（nonhierarchical cluster procedure）。这样的方法，是在一开始分类的时候，就已经预设分群个数，并根据整体的样本分布情况，预设好各聚类的中心点，然后再开始聚类分析。这种聚类方法称为 k-mean 聚类。

在图 10-2（b）图中，可以观察聚类分析如何依据距离，逐步合并个别的数据而成为聚

类的整个详细流程。在这里，发现最先被合并的是#3 与#11。在图 10-2（a）图样本分布当中，这两个数据其实是完全重叠的，理所当然就应该最优先合并。接下来，是#3、#11、#9，#1、#7、#8 与#12、#13、#15 这三大族群的合并，这是因为它们在图形上彼此的距离是一样的。这样逐步合并，最后得到以#6、#14 所形成的一个小聚类，以及其他申请人的集合所形成的另一个大的两组聚类。这就是聚类分析从个别独立的 *n* 个样本点开始，逐步合并成为最后两个大聚类的过程。

但是聚类分析是一种非参数方法，无法使用任何统计方法来判断最优的聚类个数。在实务上，可根据事实的样本数据、合并距离的长短差异，或者分析者的经验来作判断。

当然，对于样本的聚类（cluster on observations）也可以转换为对变量的聚类（cluster on variables），从而得到如图 10-3 所示的结果。

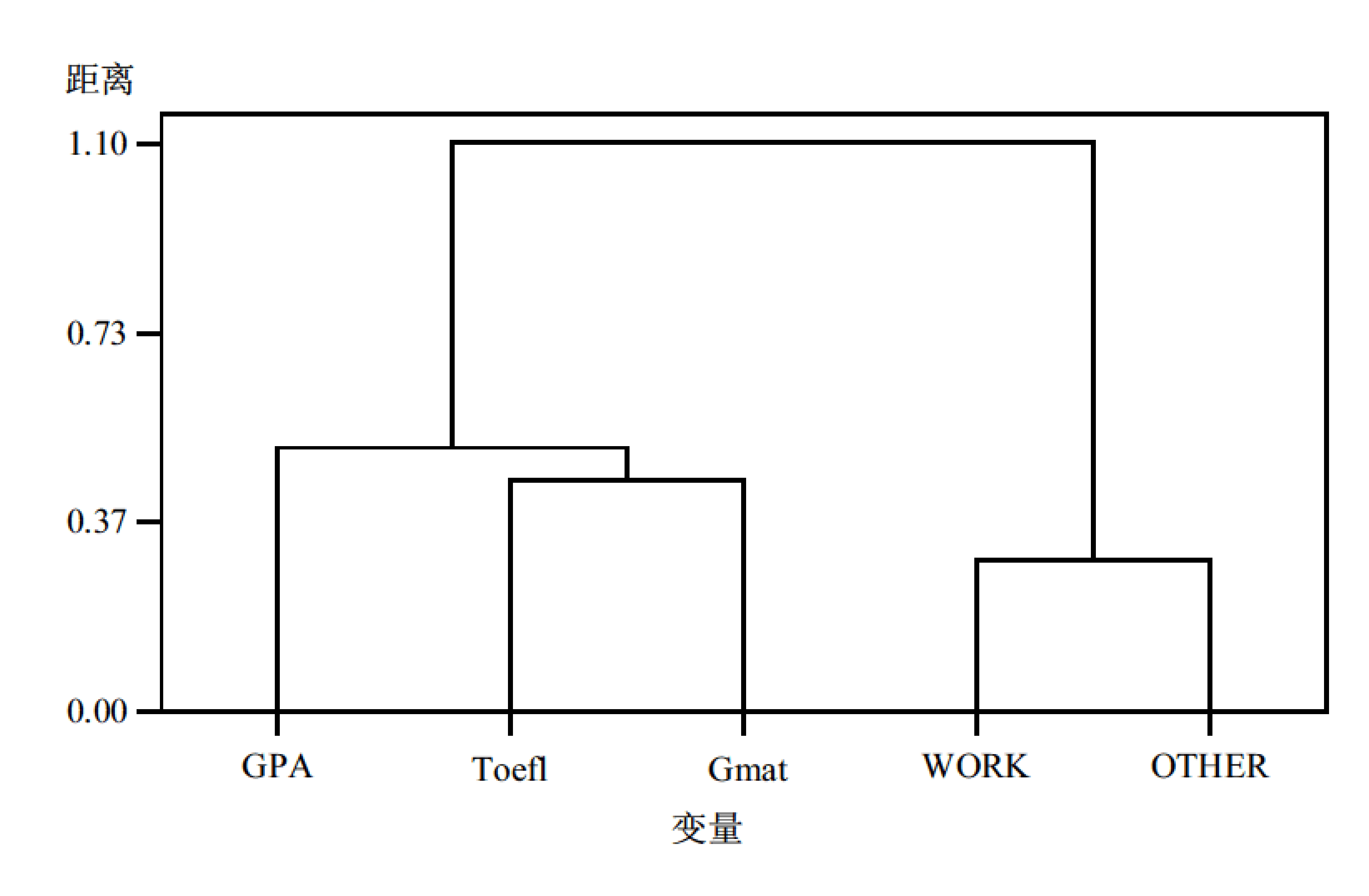


图 10-3　使用申请人数据对于不同评量标准的聚类分析结果

这时候，会发现工作经历（WORK）与其他条件（OTHER）是距离最近、最先受到合并的两个变量。综合而言，可以归纳出 WORK、OTHER 这一聚类与 GPA、Toefl、Gmat 这一聚类间可能在数据上颇有差异。

当然有一点值得注意，Gmat 与 Toefl 的计分单位比其他计分分数要高出 100 倍左右。于是在几何距离的图形衡量上，如果不注意单位问题，这两个变量便会显著超越其他变量，而错误地合并 GPA、WORK 与 OTHER 这三个变量。这时，应先将数据标准化，再做聚类分析。

10.4　Excel 2007 聚类分析

Step1：数据来源为 Microsoft 内建数据集，为 2002—2007 年自行车购买的数据集，建立聚类模型。选择【数据挖掘】→【聚类分析】命令，开始建立数据挖掘模型，弹出如图 10-4 所示的【聚类分析向导入门】窗口，然后单击【下一步】按钮。



图 10-4　【聚类分析向导入门】窗口

Step2：在如图 10-5 所示的【选择源数据】窗口的【表】下拉列表框中选择 Excel 中要分析的数据表，单击【下一步】按钮。



图 10-5　【选择源数据】窗口

Step3：在选择数据列的步骤时，选择进入聚类的变量，由于 ID 为顾客编码，所以本

次分析不将它归为进入聚类变量，取消选中 ID 前的复选框，接着单击【下一步】按钮，如图 10-6 所示。

Step4：选择聚类变量后，在【段数】栏内选择聚类个数，可以使用软件自动检测，或自行指定目标值，这里先将目标值设为 10，单击【下一步】按钮，如图 10-7 所示。



图 10-6 【聚类分析】窗口



图 10-7 选择聚类个数

Step5：完成数据挖掘模型，选中【启用钻取】复选框，然后单击【完成】按钮，如图 10-8 所示。

Step6：产生 10 个聚类的聚类图表，若将图形复制到 Excel 中再进行操作，可以单击【复制到 Excel】按钮，如图 10-9 所示。



图 10-8　单击【完成】按钮

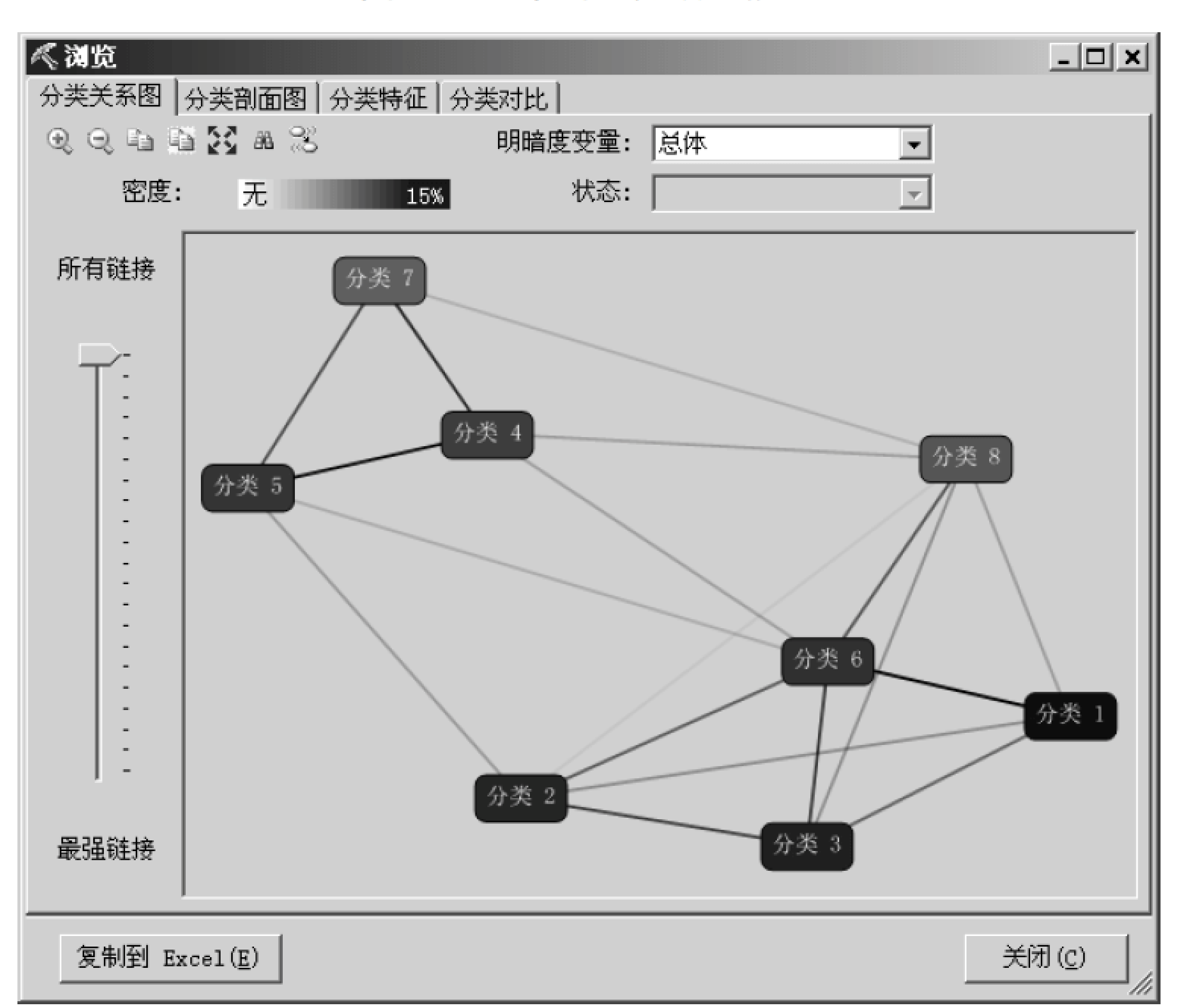


图 10-9　聚类图表

Step7：将聚类图表复制到 Excel 中，如图 10-10 所示。

Step8：选择【分类剖面图】选项卡，显示各群体在不同变量下的差异，如图 10-11 所示。单击【复制到 Excel】按钮。

Step9：将分类剖面图复制到 Excel，如图 10-12 所示。

Step10：选择【分类特征】选项卡，显示各聚类在不同变量的水平下，个体归入此类的概率值，如图 10-13 所示。

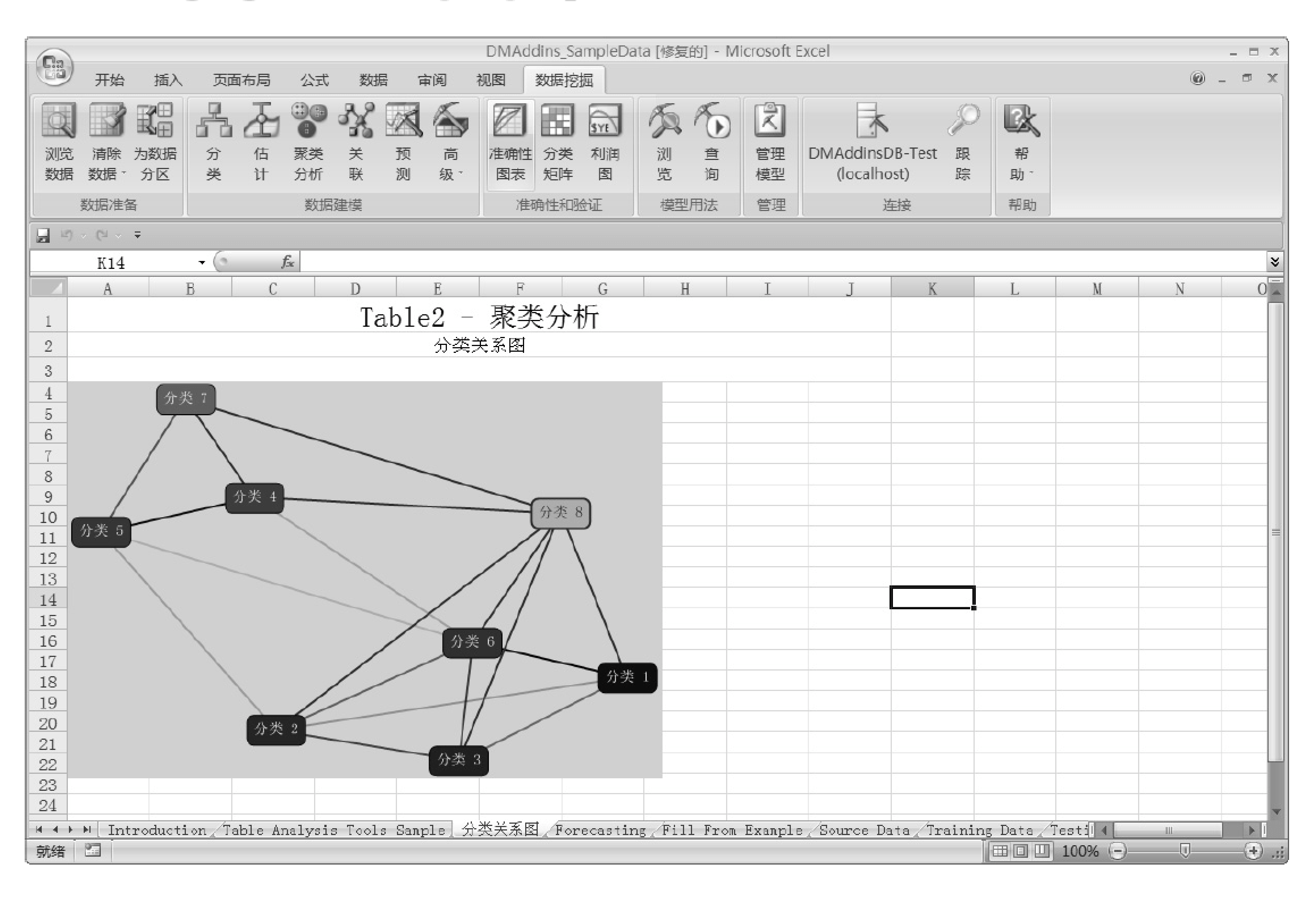


图 10-10　复制到 Excel



图 10-11　【分类剖面图】选项卡

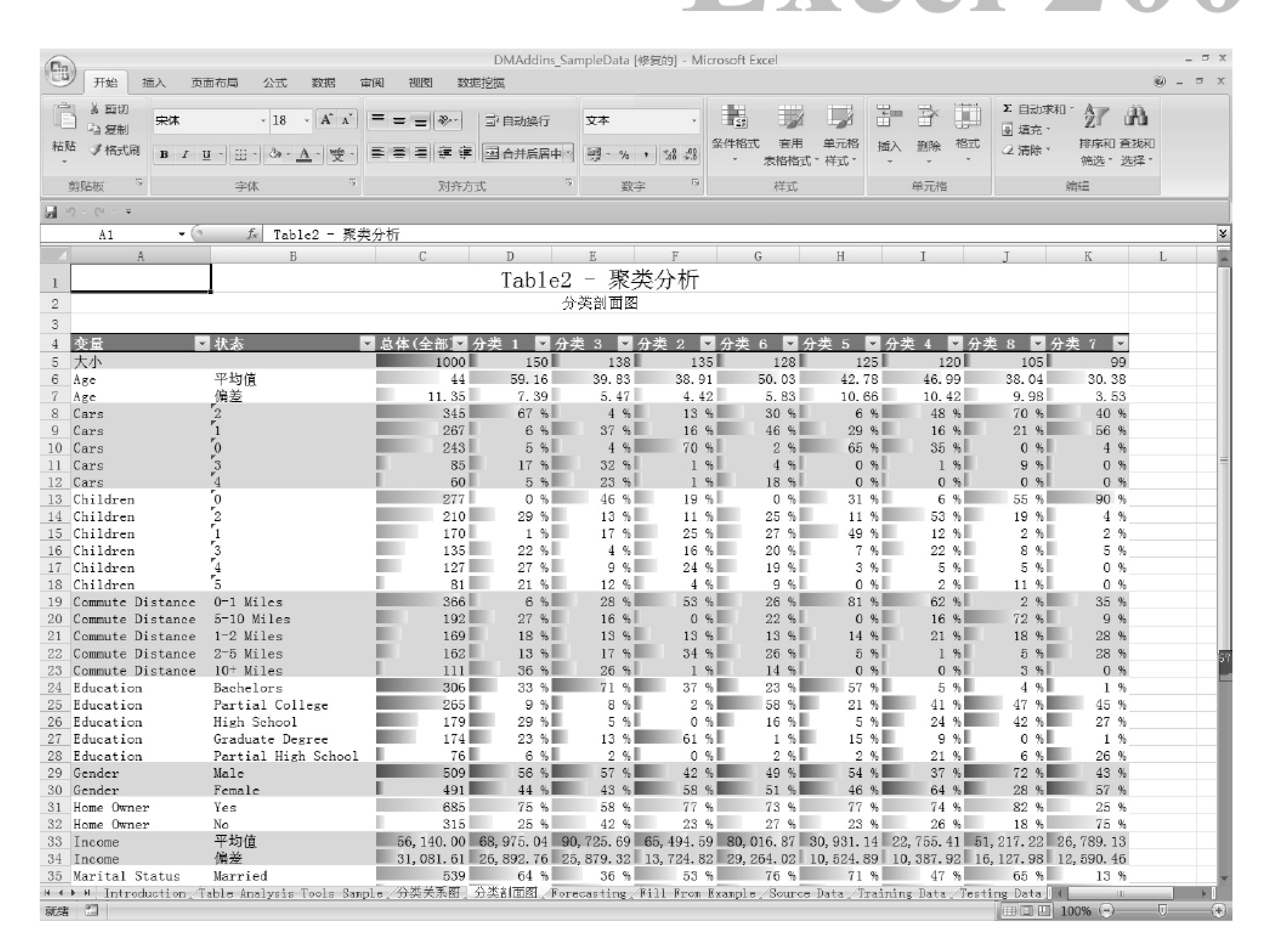
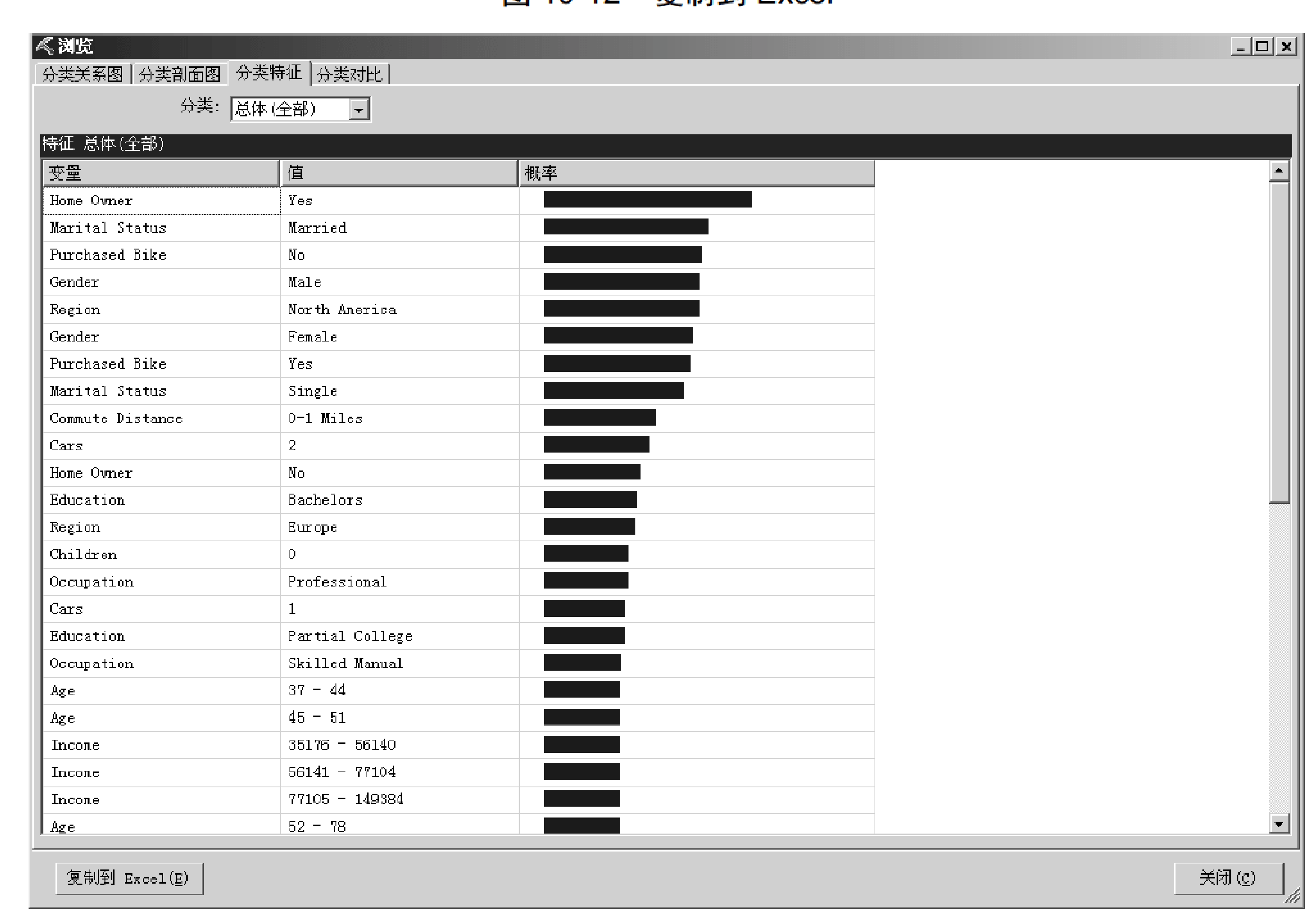


Table2 - 聚类分析

分类剖面图

变量	状态	总体(全部)	分类 1	分类 3	分类 2	分类 6	分类 5	分类 4	分类 8	分类 7
大小		1000	150	138	135	128	125	120	105	99
Age	平均值	44	59.16	39.83	38.91	50.03	42.78	46.99	38.04	30.38
Age	偏差	11.35	7.39	5.47	4.42	5.83	10.66	10.42	9.98	3.53
Cars	2	345	67 %	4 %	13 %	30 %	6 %	48 %	70 %	40 %
Cars	1	267	6 %	37 %	16 %	46 %	29 %	16 %	21 %	56 %
Cars	0	243	5 %	4 %	70 %	2 %	65 %	35 %	0 %	4 %
Cars	3	85	17 %	32 %	1 %	4 %	0 %	1 %	9 %	0 %
Cars	4	60	5 %	23 %	1 %	18 %	0 %	0 %	0 %	0 %
Children	0	277	0 %	46 %	19 %	0 %	31 %	6 %	55 %	90 %
Children	2	210	29 %	13 %	11 %	25 %	11 %	53 %	19 %	4 %
Children	1	170	1 %	17 %	25 %	27 %	49 %	12 %	2 %	2 %
Children	3	135	22 %	4 %	16 %	20 %	7 %	22 %	8 %	5 %
Children	4	127	27 %	9 %	24 %	19 %	3 %	5 %	5 %	0 %
Children	5	81	21 %	12 %	4 %	9 %	0 %	2 %	11 %	0 %
Commute Distance	0-1 Miles	366	6 %	28 %	53 %	26 %	81 %	62 %	2 %	35 %
Commute Distance	5-10 Miles	192	27 %	16 %	0 %	22 %	0 %	16 %	72 %	9 %
Commute Distance	1-2 Miles	169	18 %	13 %	13 %	13 %	14 %	21 %	18 %	28 %
Commute Distance	2-5 Miles	162	13 %	17 %	34 %	26 %	5 %	1 %	5 %	28 %
Commute Distance	10+ Miles	111	36 %	26 %	1 %	14 %	0 %	0 %	3 %	0 %
Education	Bachelors	306	33 %	71 %	37 %	23 %	57 %	5 %	4 %	1 %
Education	Partial College	265	9 %	8 %	2 %	58 %	21 %	41 %	47 %	45 %
Education	High School	179	29 %	5 %	0 %	16 %	5 %	24 %	42 %	27 %
Education	Graduate Degree	174	23 %	13 %	61 %	1 %	15 %	9 %	0 %	1 %
Education	Partial High School	76	6 %	2 %	0 %	2 %	2 %	21 %	6 %	26 %
Gender	Male	509	56 %	57 %	42 %	49 %	54 %	37 %	72 %	43 %
Gender	Female	491	44 %	43 %	58 %	51 %	46 %	64 %	28 %	57 %
Home Owner	Yes	685	75 %	58 %	77 %	73 %	77 %	74 %	82 %	25 %
Home Owner	No	315	25 %	42 %	23 %	27 %	23 %	26 %	18 %	75 %
Income	平均值	56,140.00	68,975.04	90,725.69	65,494.59	80,016.87	30,931.14	22,755.41	51,217.22	26,789.13
Income	偏差	31,081.61	26,892.76	25,879.32	13,724.82	29,264.02	10,524.89	10,387.92	16,127.98	12,590.46
Marital Status	Married	539	64 %	36 %	53 %	76 %	71 %	47 %	65 %	13 %

图 10-12 复制到 Excel



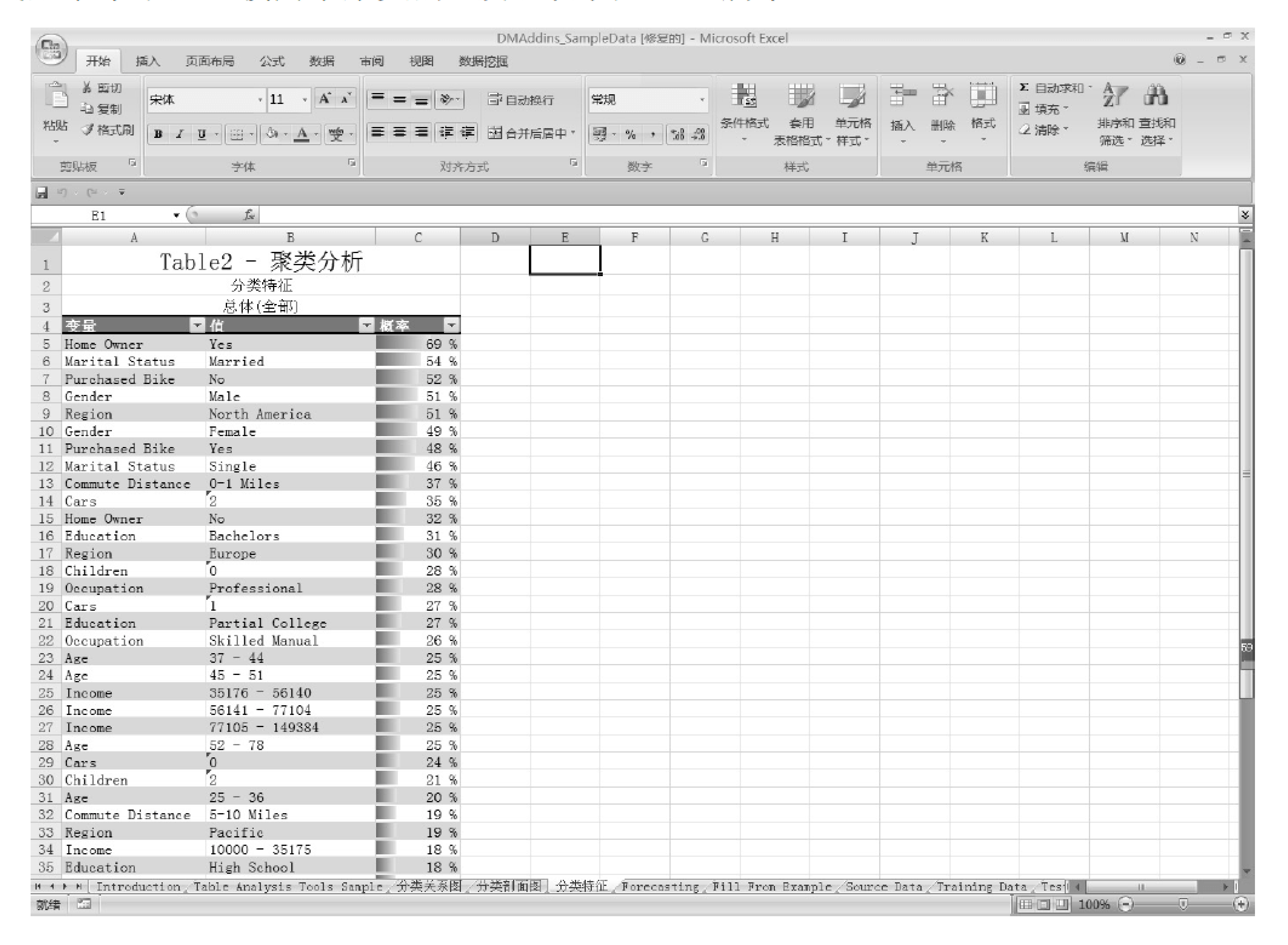
变量	值	概率
Home Owner	Yes	
Marital Status	Married	
Purchased Bike	No	
Gender	Male	
Region	North America	
Gender	Female	
Purchased Bike	Yes	
Marital Status	Single	
Commute Distance	0-1 Miles	
Cars	2	
Home Owner	No	
Education	Bachelors	
Region	Europe	
Children	0	
Occupation	Professional	
Cars	1	
Education	Partial College	
Occupation	Skilled Manual	
Age	37 - 44	
Age	45 - 51	
Income	35176 - 56140	
Income	56141 - 77104	
Income	77105 - 149384	
Age	52 - 78	

图 10-13 【分类特征】选项卡

Step11：将图表复制到 Excel，如图 10-14 所示。

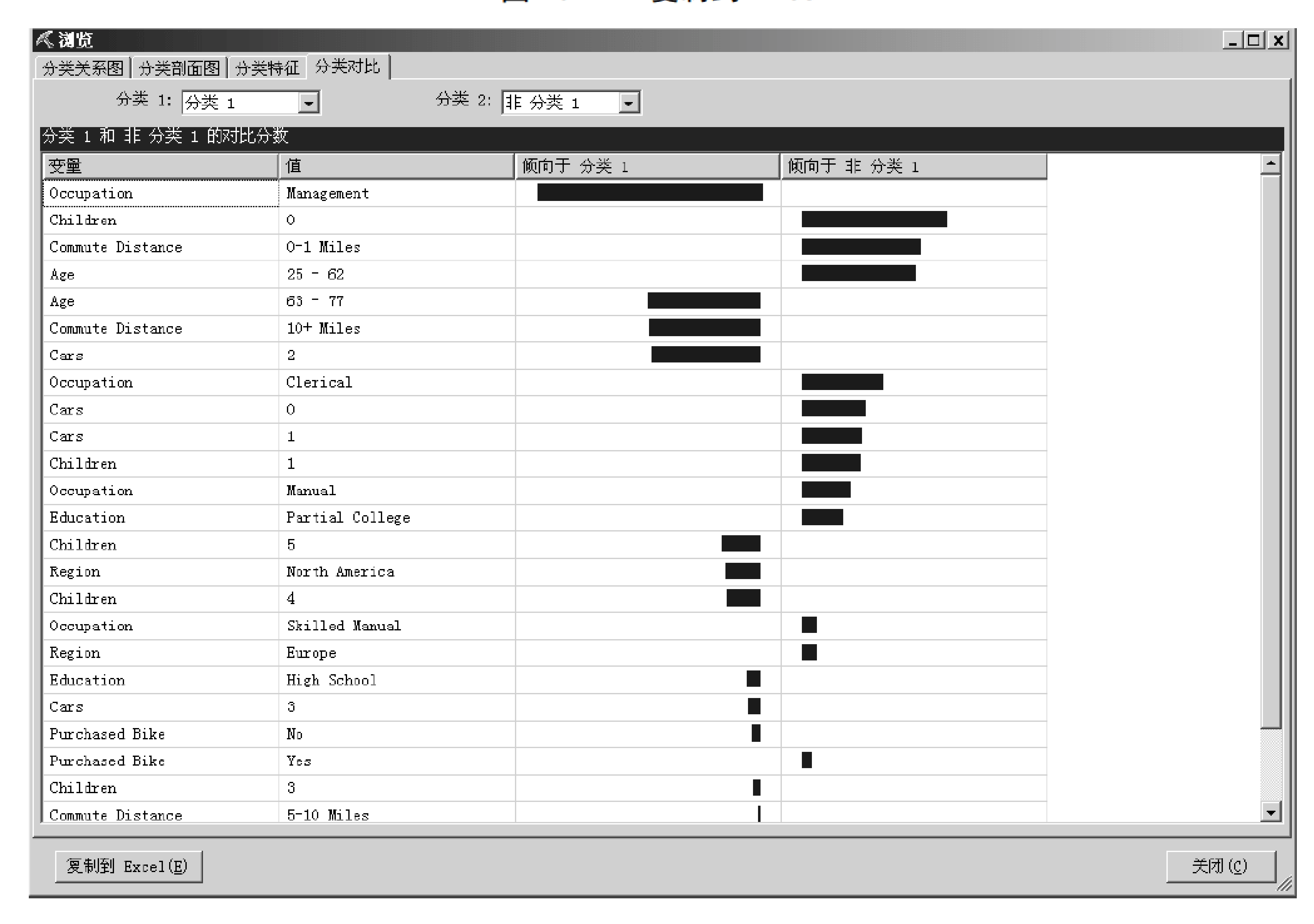
Step12：选择【分类对比】选项卡，可以在图形上方选择要比较的两个聚类，在不同

的变量水平下，比较两个聚类的差异，如图 10-15 所示。



变量	值	概率
Home Owner	Yes	69 %
Marital Status	Married	54 %
Purchased Bike	No	52 %
Gender	Male	51 %
Region	North America	51 %
Gender	Female	49 %
Purchased Bike	Yes	48 %
Marital Status	Single	46 %
Commute Distance	0-1 Miles	37 %
Cars	2	35 %
Home Owner	No	32 %
Education	Bachelors	31 %
Region	Europe	30 %
Children	0	28 %
Occupation	Professional	28 %
Cars	1	27 %
Education	Partial College	27 %
Occupation	Skilled Manual	26 %
Age	37 - 44	25 %
Age	45 - 51	25 %
Income	35176 - 56140	25 %
Income	56141 - 77104	25 %
Income	77105 - 149384	25 %
Age	52 - 78	25 %
Cars	0	24 %
Children	2	21 %
Age	25 - 36	20 %
Commute Distance	5-10 Miles	19 %
Region	Pacific	19 %
Income	10000 - 35175	18 %
Education	High School	18 %

图 10-14　复制到 Excel



变量	值	倾向于 分类 1	倾向于 非 分类 1
Occupation	Management		
Children	0		
Commute Distance	0-1 Miles		
Age	25 - 62		
Age	63 - 77		
Commute Distance	10+ Miles		
Cars	2		
Occupation	Clerical		
Cars	0		
Cars	1		
Children	1		
Occupation	Manual		
Education	Partial College		
Children	5		
Region	North America		
Children	4		
Occupation	Skilled Manual		
Region	Europe		
Education	High School		
Cars	3		
Purchased Bike	No		
Purchased Bike	Yes		
Children	3		
Commute Distance	5-10 Miles		

图 10-15　【分类对比】选项卡

Step13：将图表复制到 Excel，如图 10-16 所示。

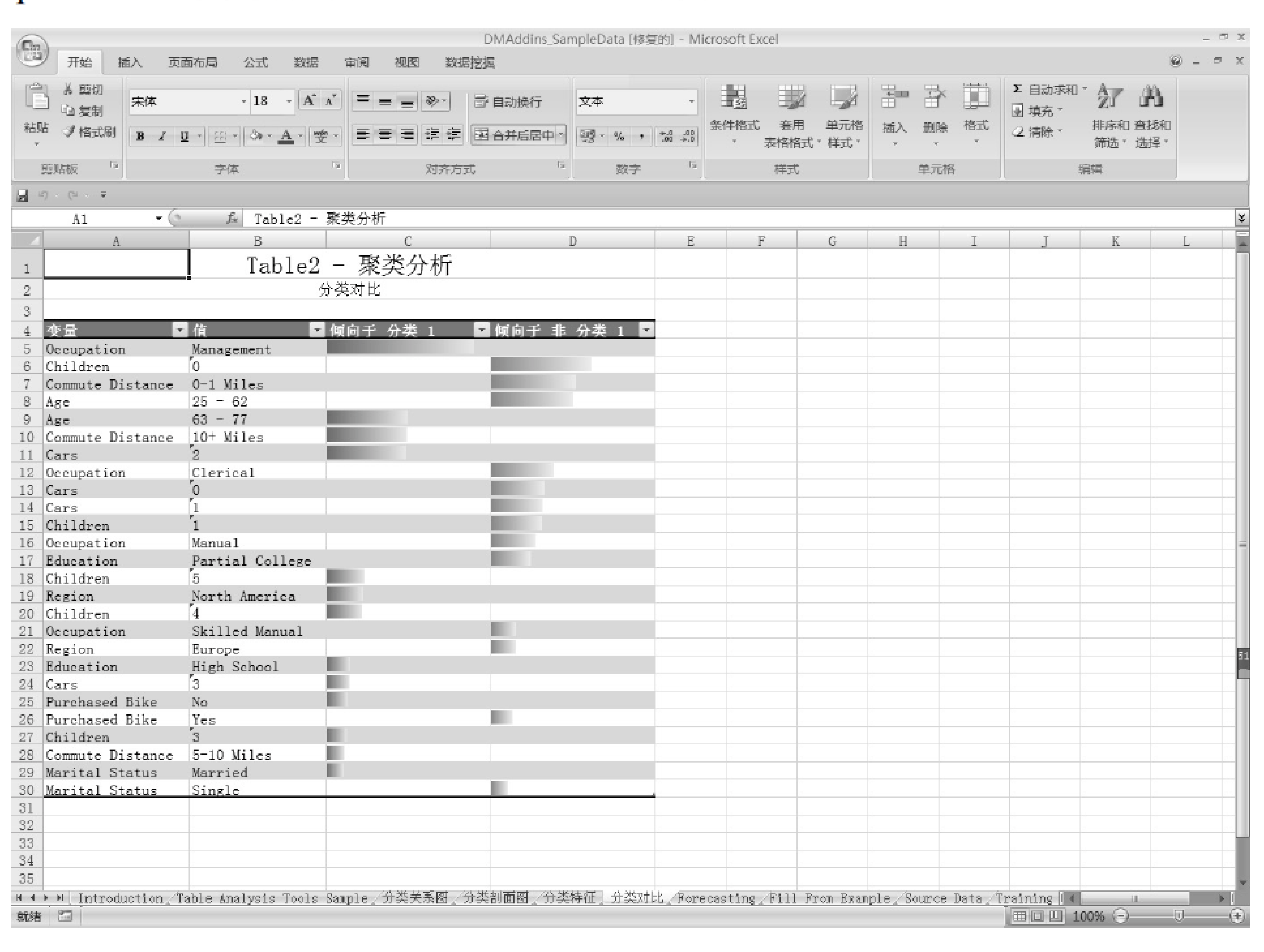


图 10-16 复制到 Excel

Step14：同样地，单击【数据挖掘】中的【高级】按钮，开始进行建立数据挖掘模型。单击【下一步】按钮，如图 10-17 所示。



图 10-17 建立数据挖掘模型

Step15：在如图 10-18 所示的【选择挖掘算法】窗口中，在【算法】下拉列表框中选择 Microsoft 聚类分析，单击【下一步】按钮。

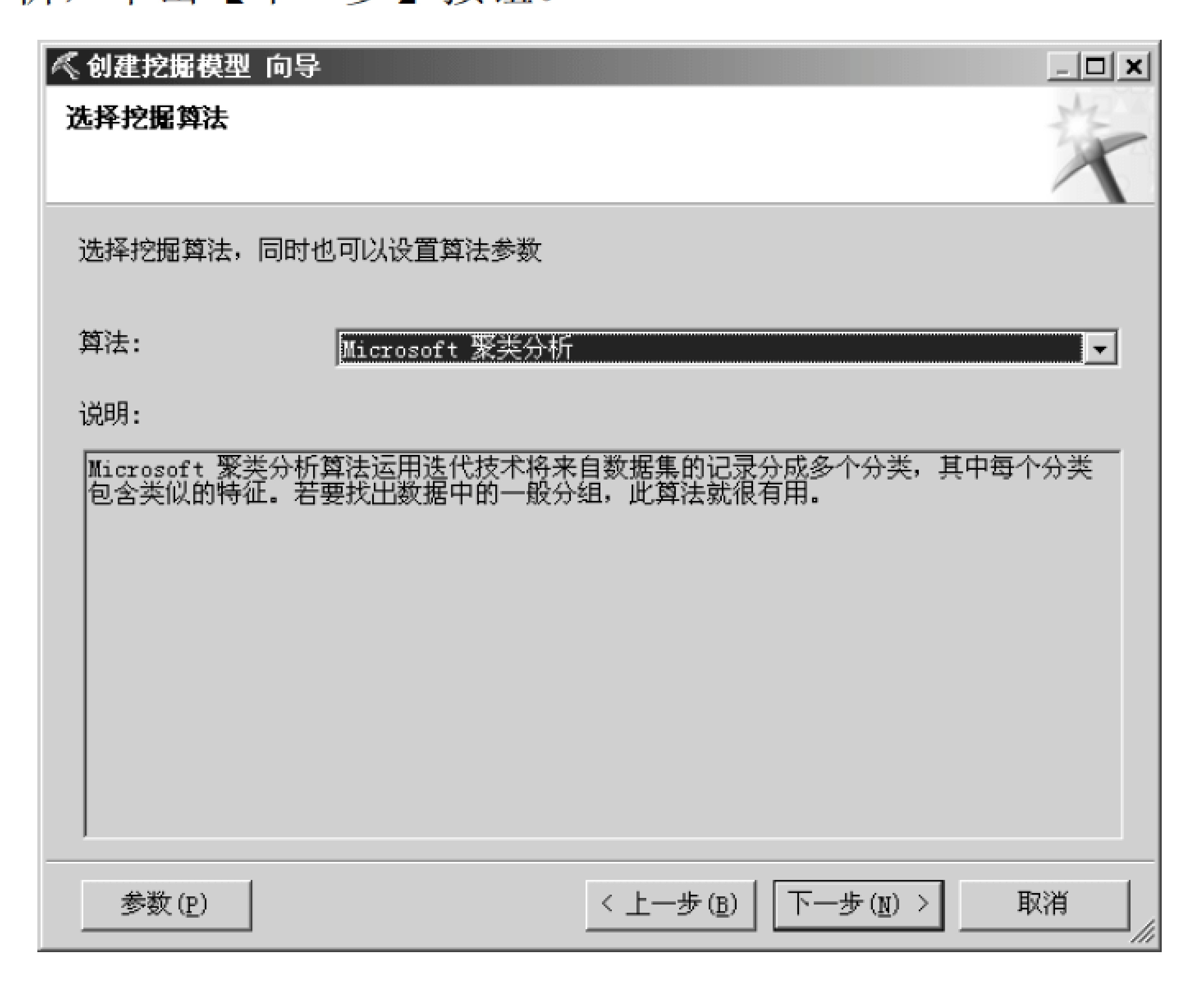


图 10-18 【选择挖掘算法】窗口

Step16：在如图 10-19 所示的【选择列】窗口中，在各个变量后方有一个下拉列表框是使用方式选取，用户可以选取各个变量的使用方式，包括“输入”、“仅预测”、“输入和预测”、“键”以及“不使用”等。本次使用是否购买自行车（purchased bike）作为预测变量 *Y*，其余变量作为解释变量建立模型，接着单击【下一步】按钮。



图 10-19 【选择列】窗口

Step17：在如图 10-20 所示的【完成】窗口中，单击【完成】按钮，软件进行建立数据挖掘模型的操作。



图 10-20 【完成】窗口

Step18：产生如图 10-21 所示的聚类图表，其余选项卡都与前文类似，不再赘述。

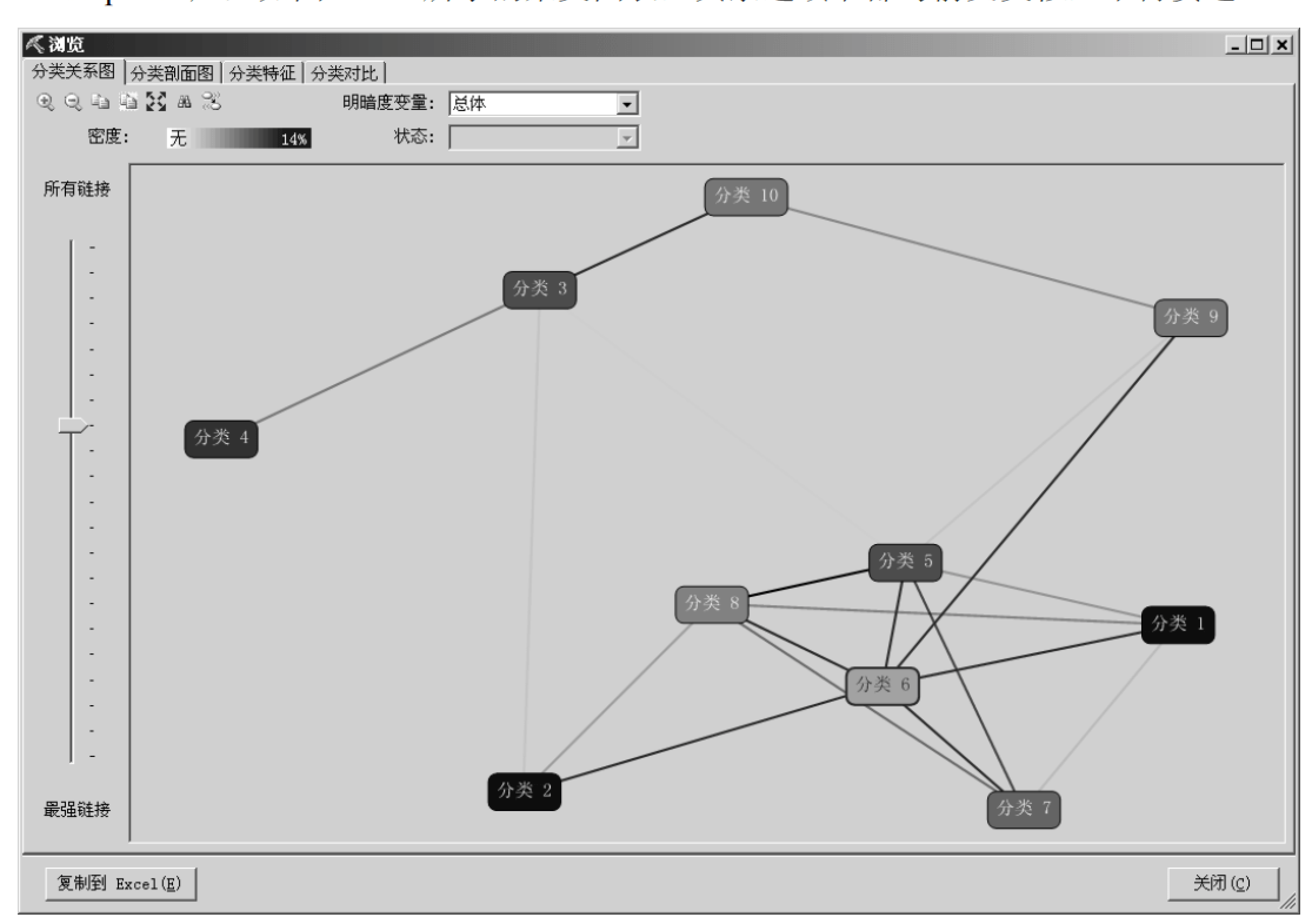


图 10-21 聚类图表

Step19：单击【数据挖掘】中的【准确性图表】按钮，弹出如图 10-22 所示的【准确性图表向导入门】窗口。接着单击【下一步】按钮。

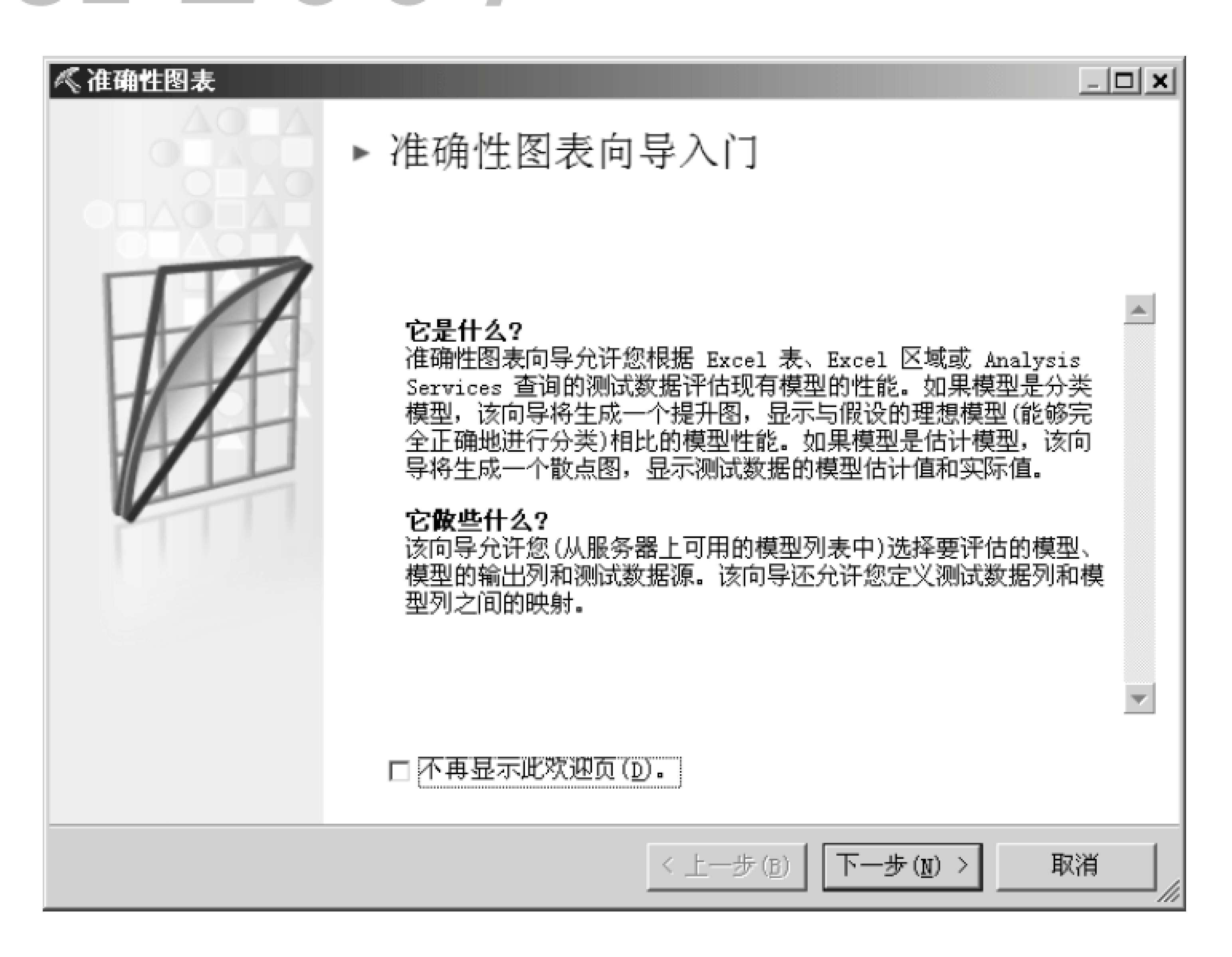


图 10-22 【准确性图表向导入门】窗口

Step20：在如图 10-23 所示的【指定要预测的列和要预测的值】窗口中，选择进行预测的数据列，本次选择自行车购买图表，单击【下一步】按钮。

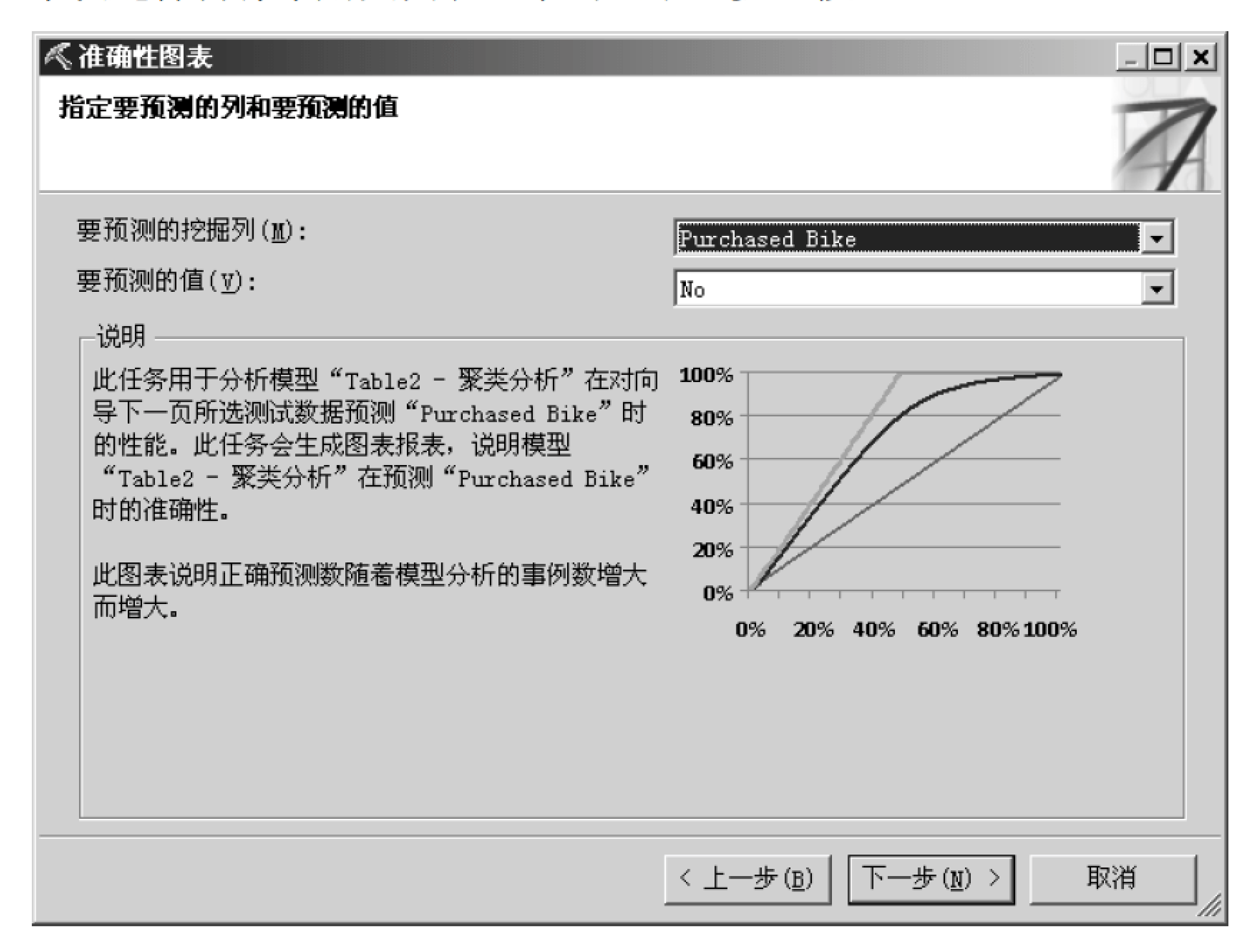


图 10-23 【指定要预测的列和要预测的值】窗口

Step21：在如图 10-24 所示的【指定关系】窗口中，选择变量间的关系，单击【完成】按钮。



图 10-24 【指定关系】窗口

Step22：将图表复制到 Excel 中，如图 10-25 所示。

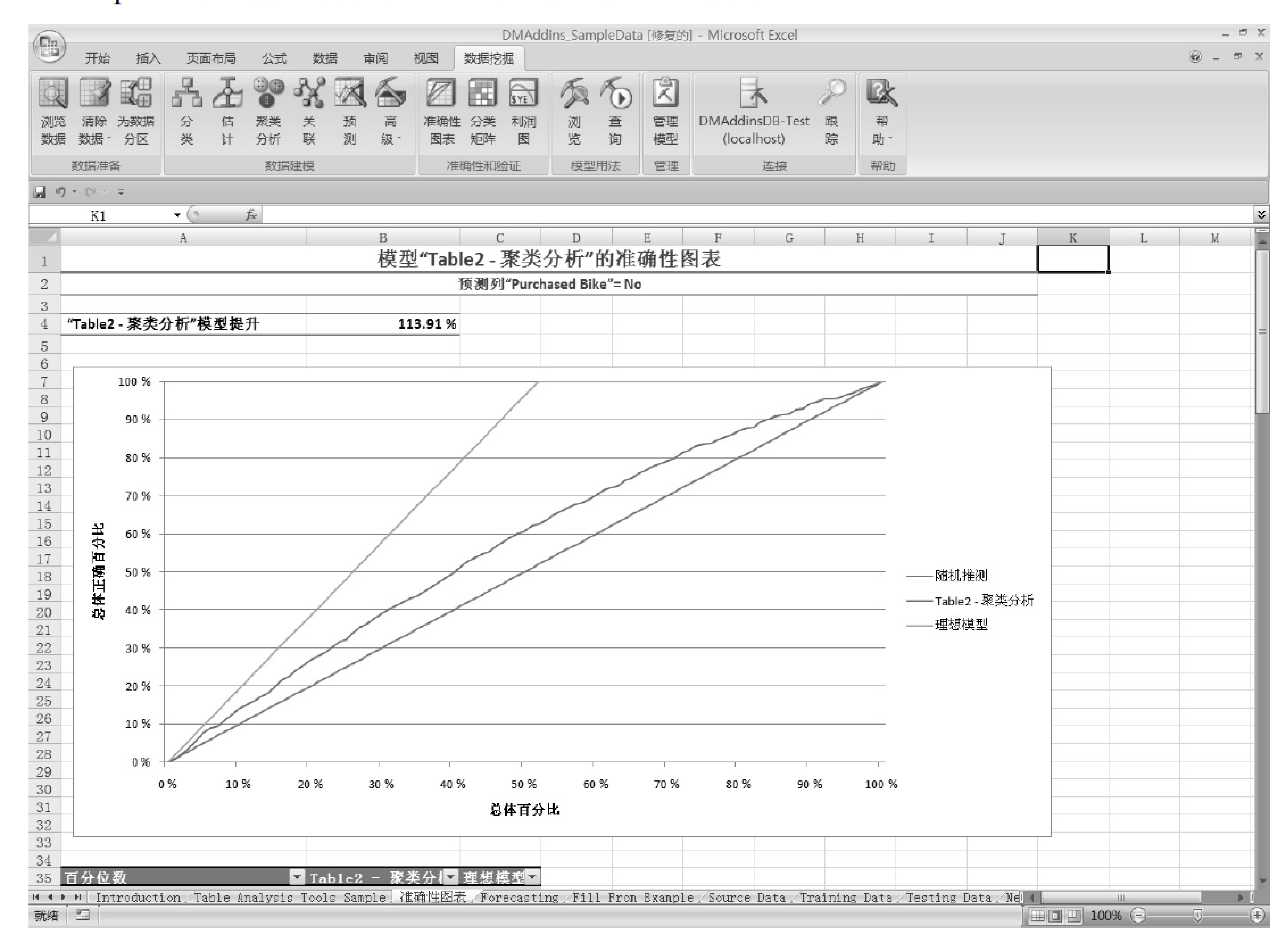


图 10-25 复制到 Excel

Step23：单击【数据挖掘】中的【分类矩阵】按钮，弹出如图 10-26 所示的【分类矩阵向导入门】窗口，接着单击【下一步】按钮。



图 10-26 【分类矩阵向导入门】窗口

Step24：在如图 10-27 所示的【指定要预测的列】窗口中，选择预测的数据列，即以自行车购买作为分析变量，单击【下一步】按钮。

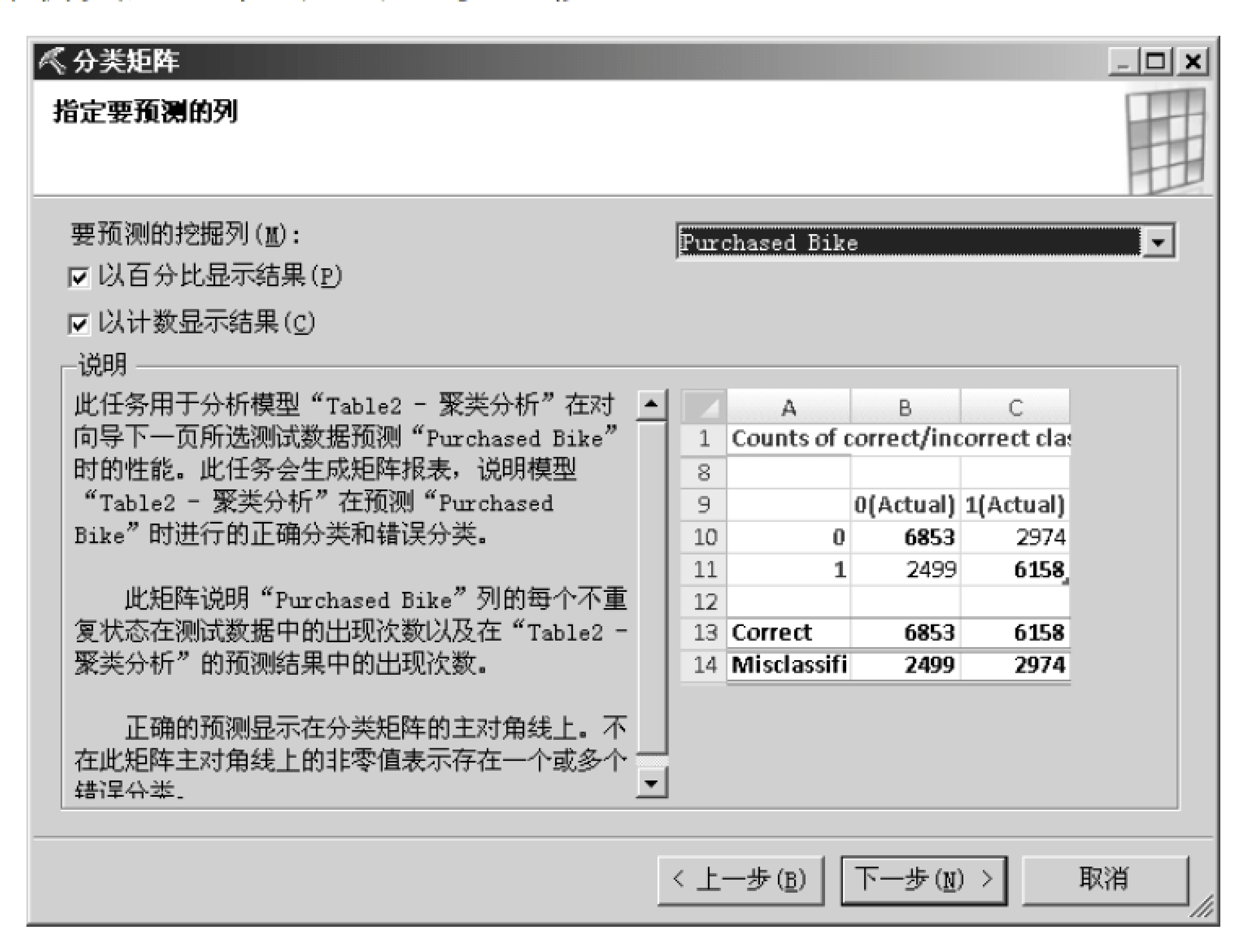


图 10-27 【指定要预测的列】窗口

Step25：在如图 10-28 所示的【指定关系】窗口中，选择变量间的关系，单击【完成】按钮。



图 10-28 【指定关系】窗口

Step26：复制分类矩阵至 Excel 中，如图 10-29 所示。

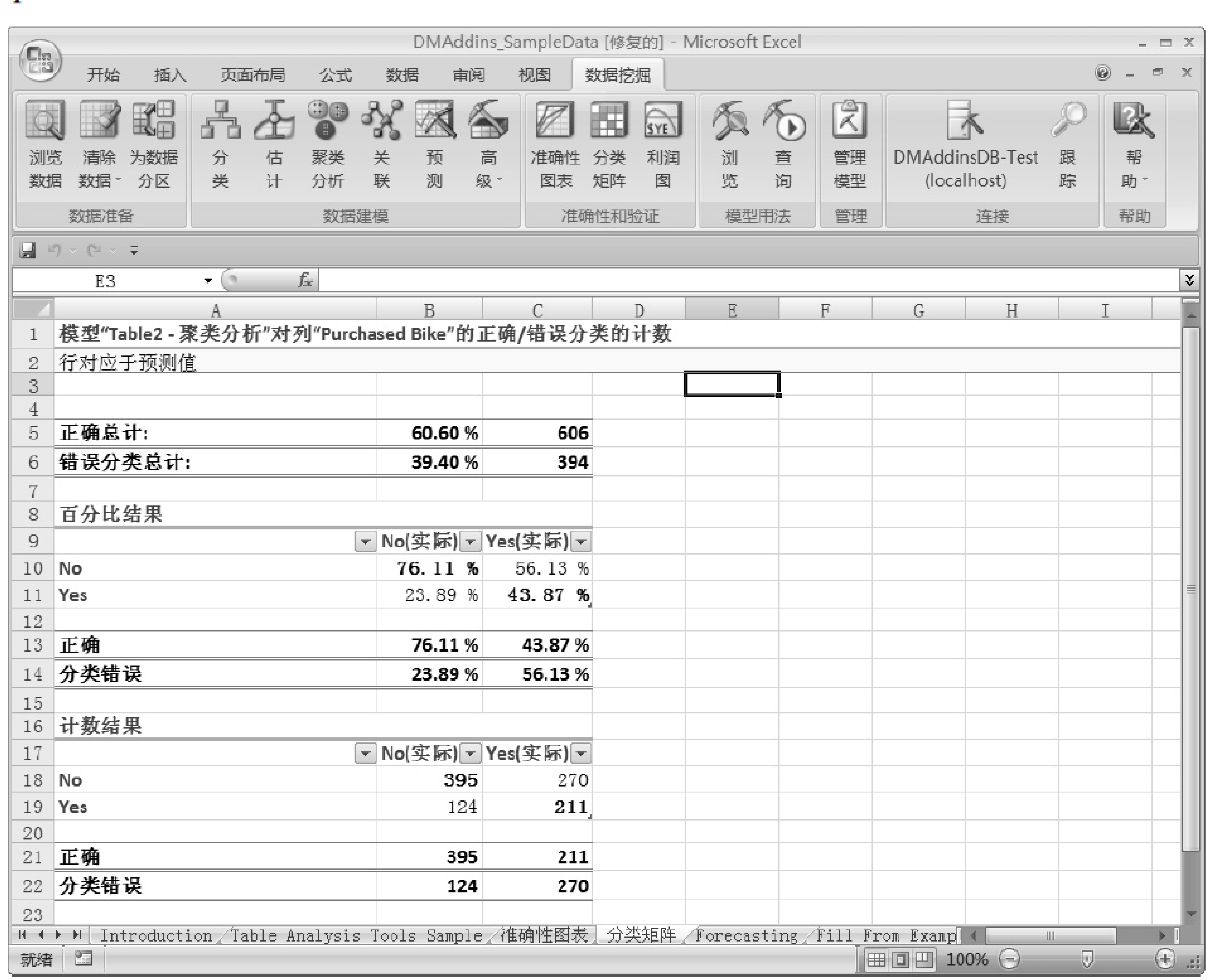


图 10-29 复制到 Excel

Step27：单击【数据挖掘】中的【利润图】按钮，弹出如图 10-30 所示的【利润图向导入门】窗口，单击【下一步】按钮。

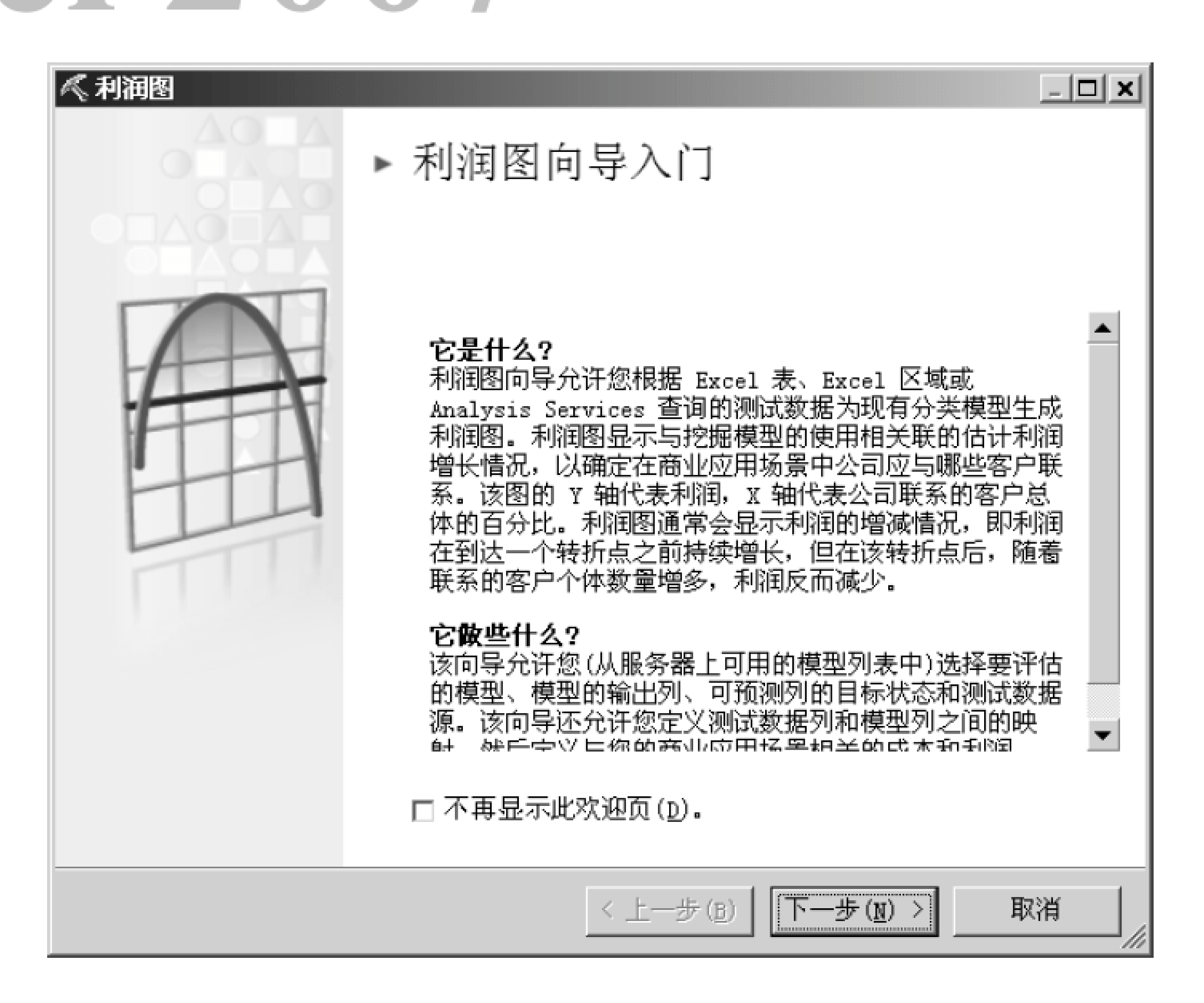


图 10-30 【利润图向导入门】窗口

Step28：在如图 10-31 所示的【指定利润图参数】窗口中，选择要分析的变量，单击【下一步】按钮。

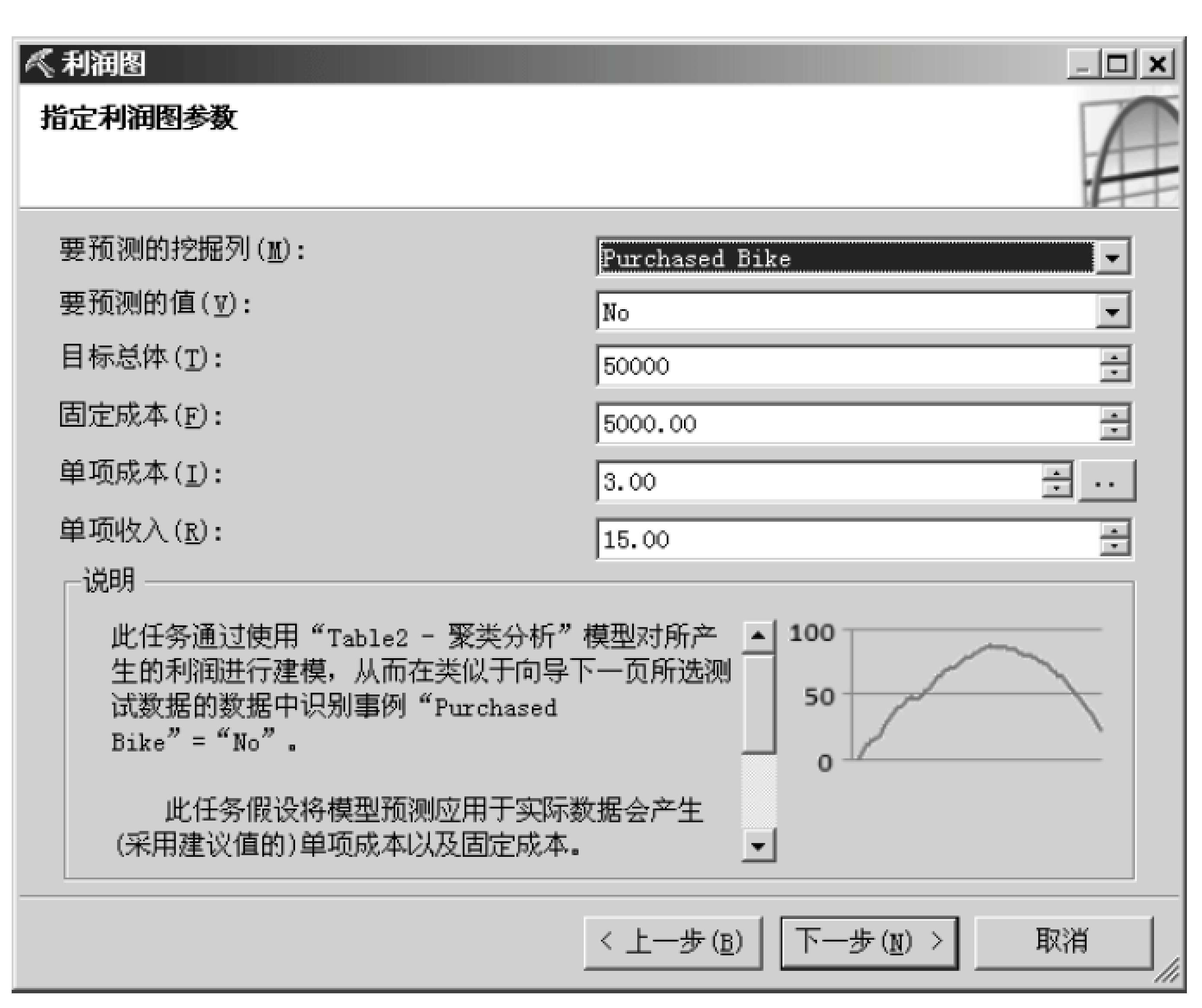


图 10-31 【指定利润图参数】窗口

Step29：在如图 10-32 所示的【指定关系】窗口中，选择变量间的关系，单击【完成】按钮。



图 10-32 【指定关系】窗口

Step30：复制利润图到 Excel 中，如图 10-33 所示。

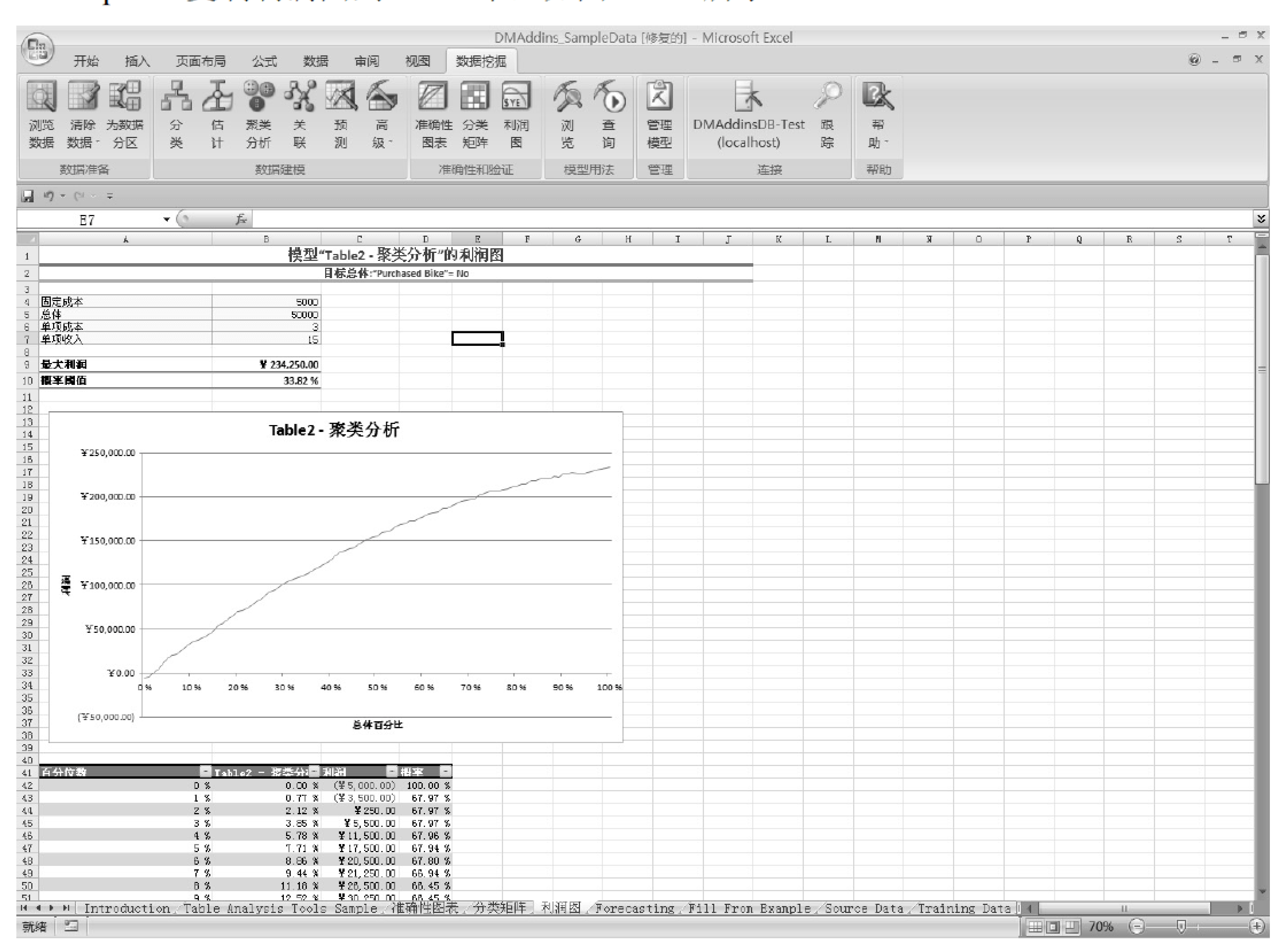


图 10-33 复制到 Excel

第 11 章　时 序 聚 类

11.1　基 本 概 念

时间序列数据是非常广泛的，例如网络用户经常按照各种链接浏览网站的点击顺序记录。时序聚类算法可以根据用户浏览的网页顺序对其进行分组，来分析网络用户的行为，并确定某些链接是否具有更高的访问率。时序聚类算法还可用于访问行为预测。利用顾客购买物品的时间序列数据可以分析顾客所购买物品和时间的相关性，有相同或类似行为的顾客会被分在相同的聚类中。这样的分析不但包含了购买物品之间的相关性，也包含了购买物品在时间上的相关性，所以对这样的数据做聚类，在应用上会更有弹性和扩充性。

11.2　相关研究和算法

时序聚类算法都是以一般的聚类算法为基础，并充分考虑了个体在时间上的行为特征。这里列举一些经典的算法，以帮助读者获得一般的认知。

BIRCH（balanced iterative reducing and clustering using hierarchies）算法应用聚类特征树（clustering feature tree）的数据结构来建立聚类的层次结构。BIRCH 可以动态地增加个体数，聚类特征树用来存放聚类的主要信息，如个体数、个体间距离的线性和与平方和。具体步骤是先扫描数据库建立聚类特征树，再利用所得的聚类特征树进行聚类，这样可以减少聚类中 I/O 的耗费，但此算法只应用于数值型数据。有序聚类中，先找出数据里序列集合（sequence sets）中共同发生的频繁模式（co-occurrence of frequent pattern），再利用 jaccard coefficient 计算数据中序列对的相似度，最后使用凝聚的层次聚类算法（agglomerative hierarchical clustering algorithm）逐渐合并，求出所要的聚类结果。但是这样的方法只能处理静态有序序列，无法处理动态有序序列。如果新加入数据时，该算法必须要重新计算。

对于含有时间间隔的有序序列，这样的数据序列可能包含数值和类别两种不同的数据类型。一般采用三种不同的相似度指标进行计算：①事件种类相似度；②事件发生周期相似度；③基于相同子序列长度的相似度进行两两序列的相似度计算，取这三种相似度的均值作为序列的相似度。但相似度不仅包含了数据间的先后关系，还考虑了事件发生的时间间隔。计算出两两之间的相似度后，以层次聚类进行合并，直到终止条件满足为止。

11.3 Excel 2007 时序聚类

Step1：单击【高级】中的【创建挖掘模型】按钮，如图 11-1 所示。

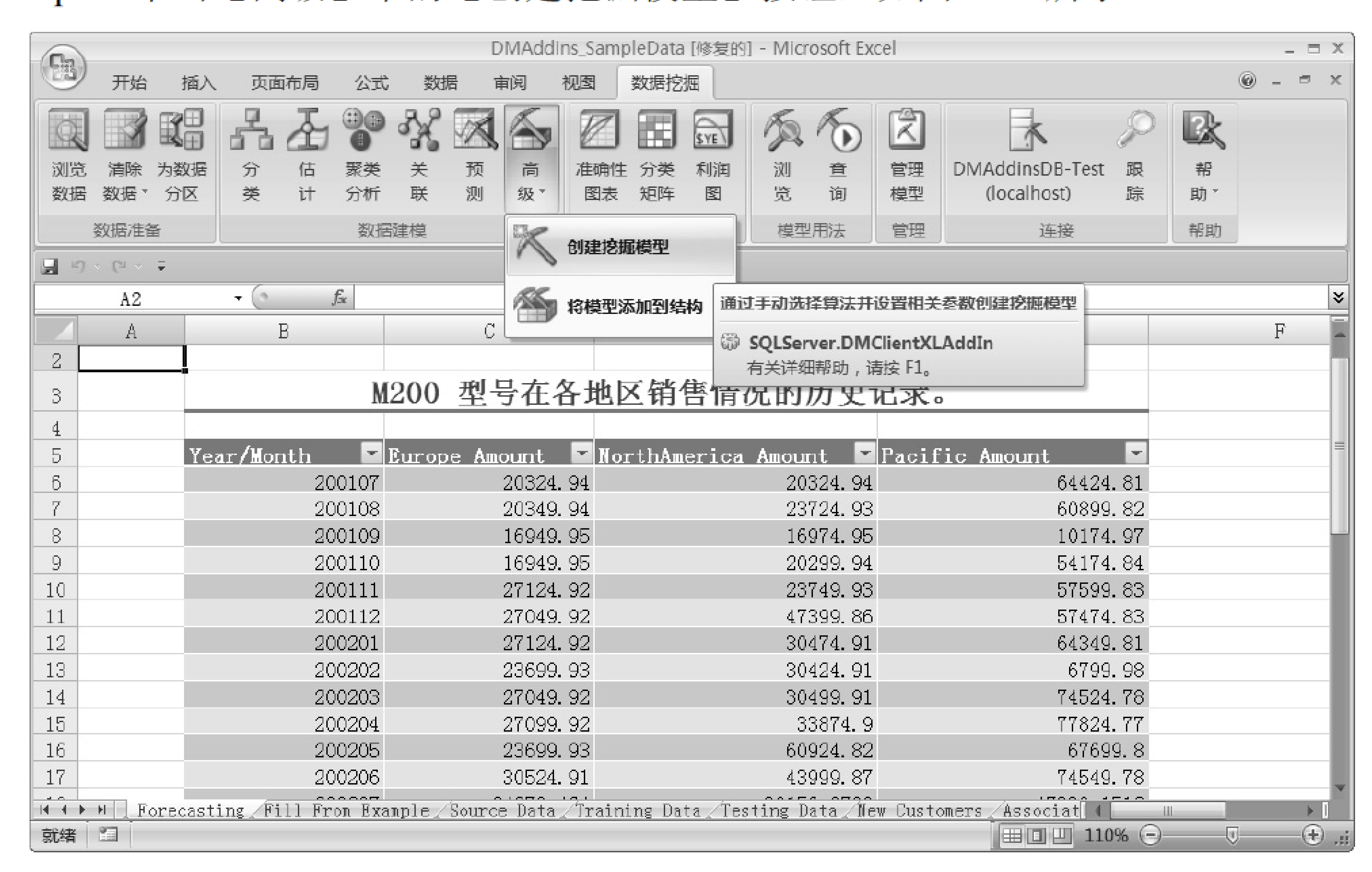


图 11-1 创建挖掘模型

Step2：弹出如图 11-2 所示的【创建模型向导入门】窗口，单击【下一步】按钮。



图 11-2 【创建模型向导入门】窗口

Step3：在如图 11-3 所示的【选择源数据】窗口中，选择数据表或者数据范围，单击【下一步】按钮。



图 11-3 【选择源数据】窗口

Step4：如图 11-4 所示的【选择挖掘算法】窗口中，在【算法】下拉列表框中选择 Microsoft 顺序分析和聚类分析，单击【下一步】按钮。



图 11-4 【选择挖掘算法】窗口

Step5：在如图 11-5 所示的【选择列】窗口中，选择变量。



图 11-5 【选择列】窗口

Step6：单击【下一步】按钮，弹出如图 11-6 所示的【完成】窗口。



图 11-6 【完成】窗口

Step7：在如图 11-7 所示的【分类特征】选项卡中显示聚类特征，将聚类分为两群。



图 11-7 【分类特征】选项卡

Step8：在如图 11-8 所示的【分类关系图】选项卡中，显示聚类特征。

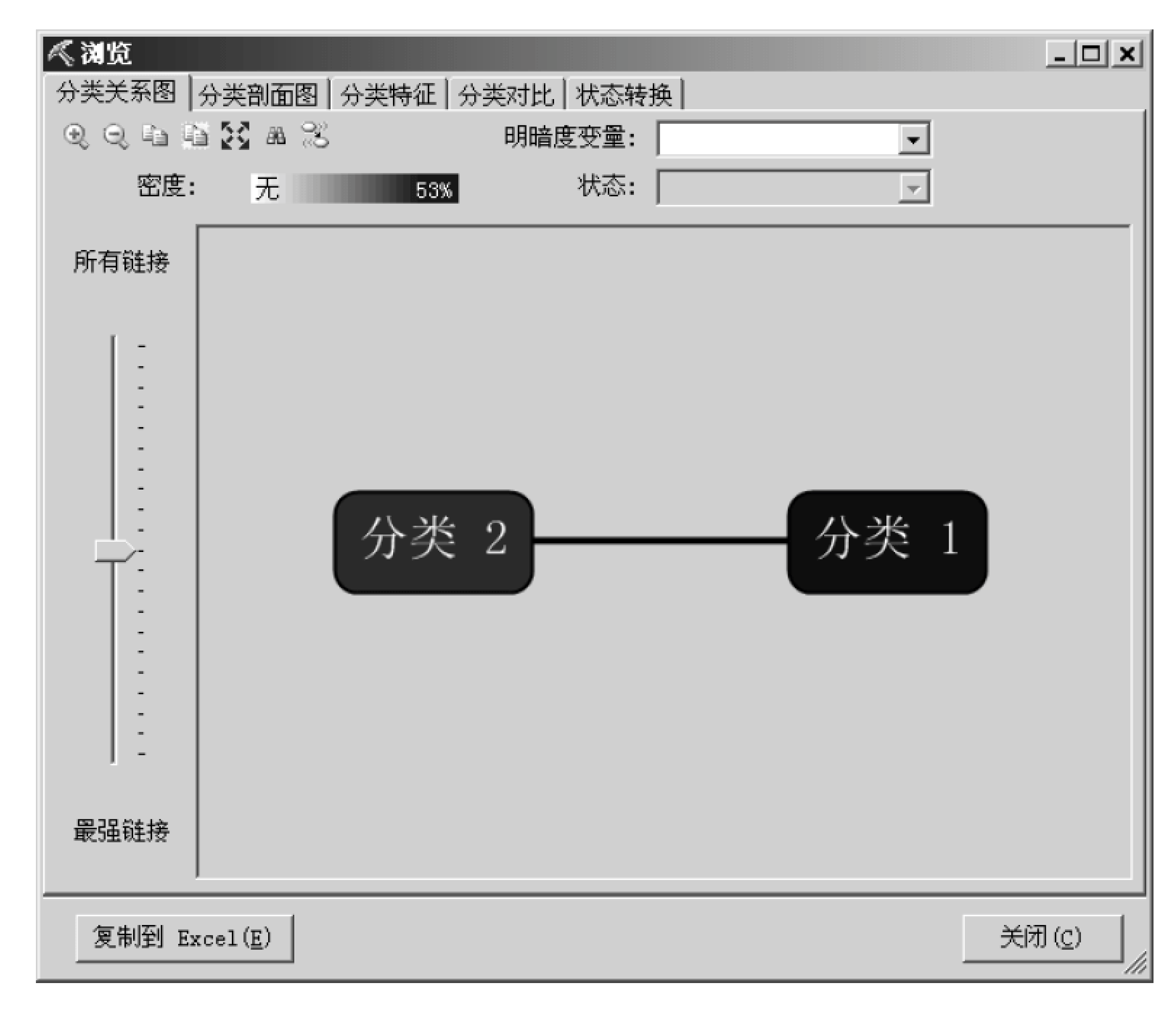


图 11-8 【分类关系图】选项卡

Step9：在如图 11-9 所示的【分类对比】选项卡中，显示聚类 2 与非聚类 2 的对比

分析。

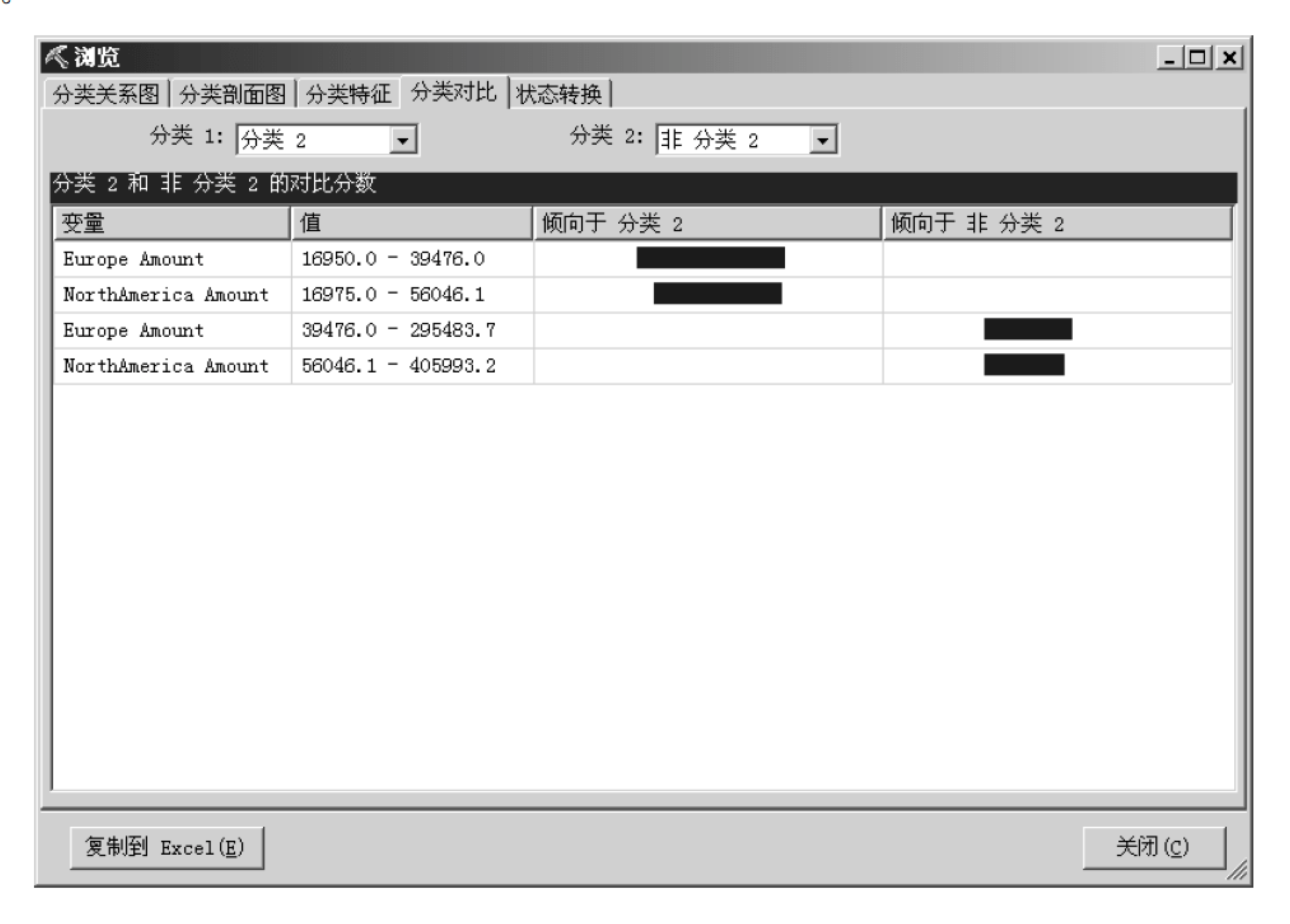


图 11-9 【分类对比】选项卡

Step10：单击【数据挖掘】中的【准确性图表】按钮，如图 11-10 所示。

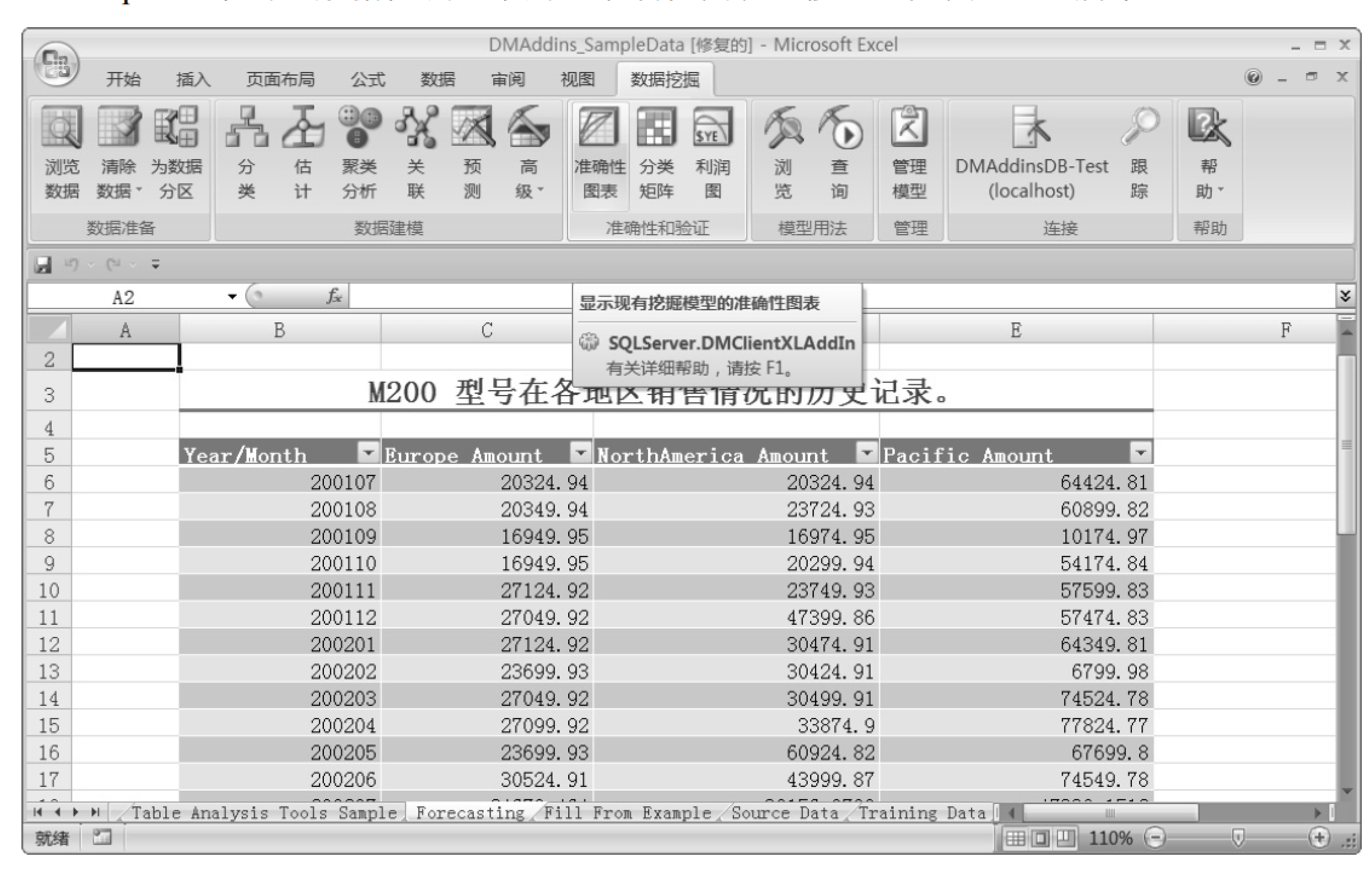


图 11-10 准确性图表

Step11：在如图 11-11 所示的【准确性图表向导入门】窗中单击【下一步】按钮。

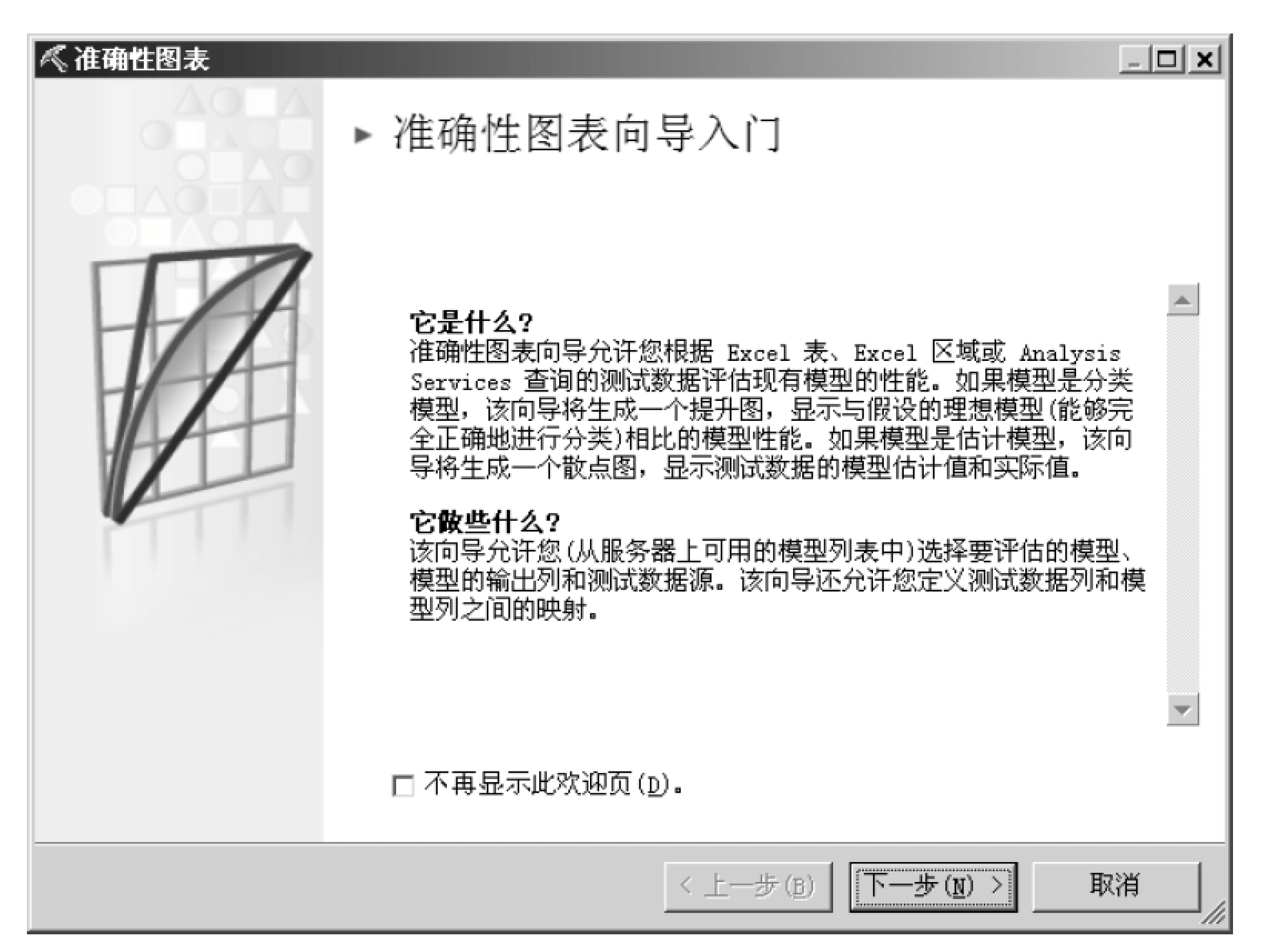


图 11-11 【准确性图表向导入门】窗口

Step12：在如图 11-12 所示的【选择模型】窗口中，单击【下一步】按钮。



图 11-12 【选择模型】窗口

Step13：在图 11-13 所示的【指定要预测的列和要预测的值】窗口中，单击【下一步】按钮。

Step14：在如图 11-14 所示的【选择源数据】窗口中，选中【表】单选按钮，并在下拉列表框中选择数据表。

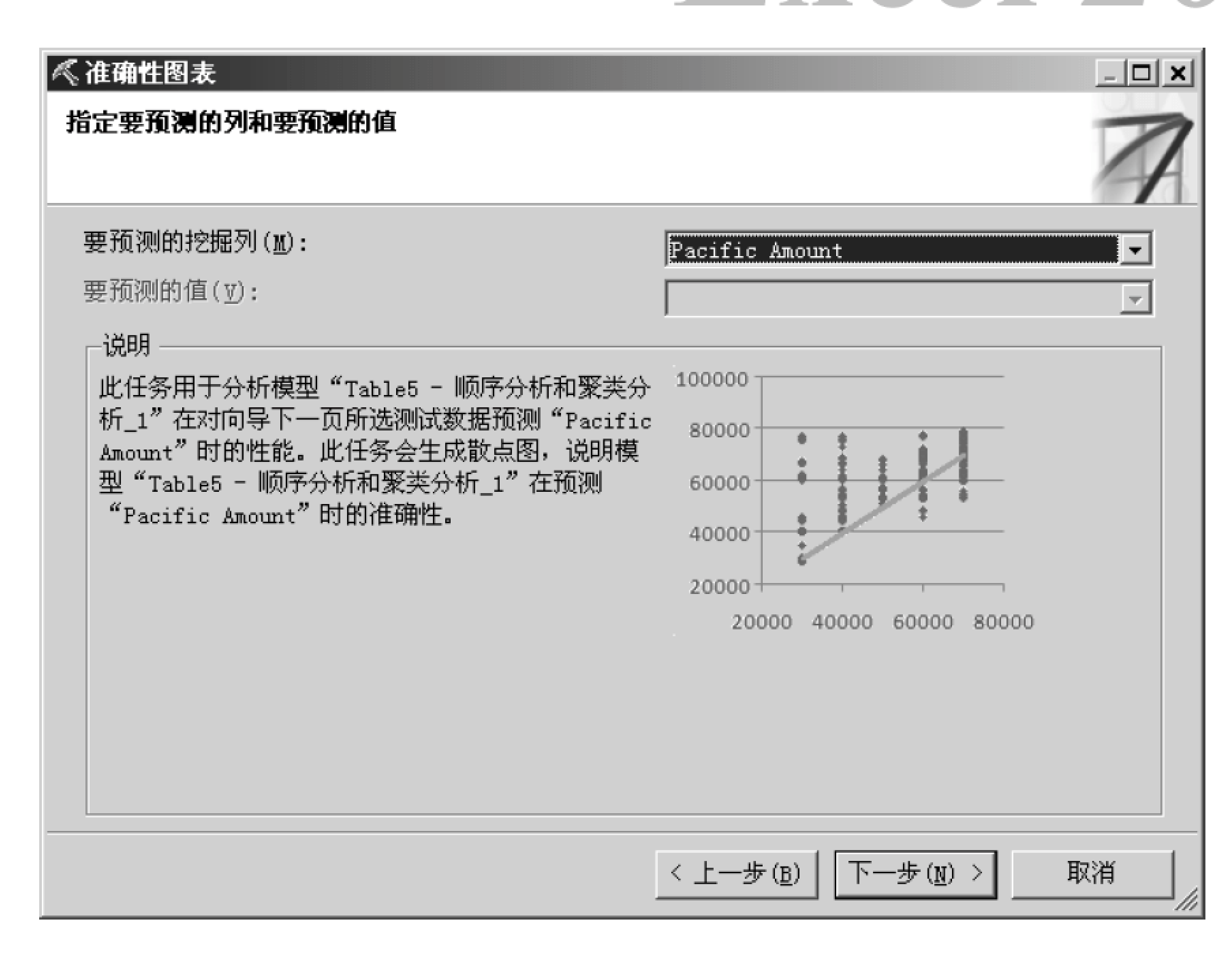


图 11-13 【指定要预测的列和要预测的值】窗口



图 11-14 【选择源数据】窗口

Step15：单击【下一步】按钮，弹出如图 11-15 所示的【指定关系】窗口。

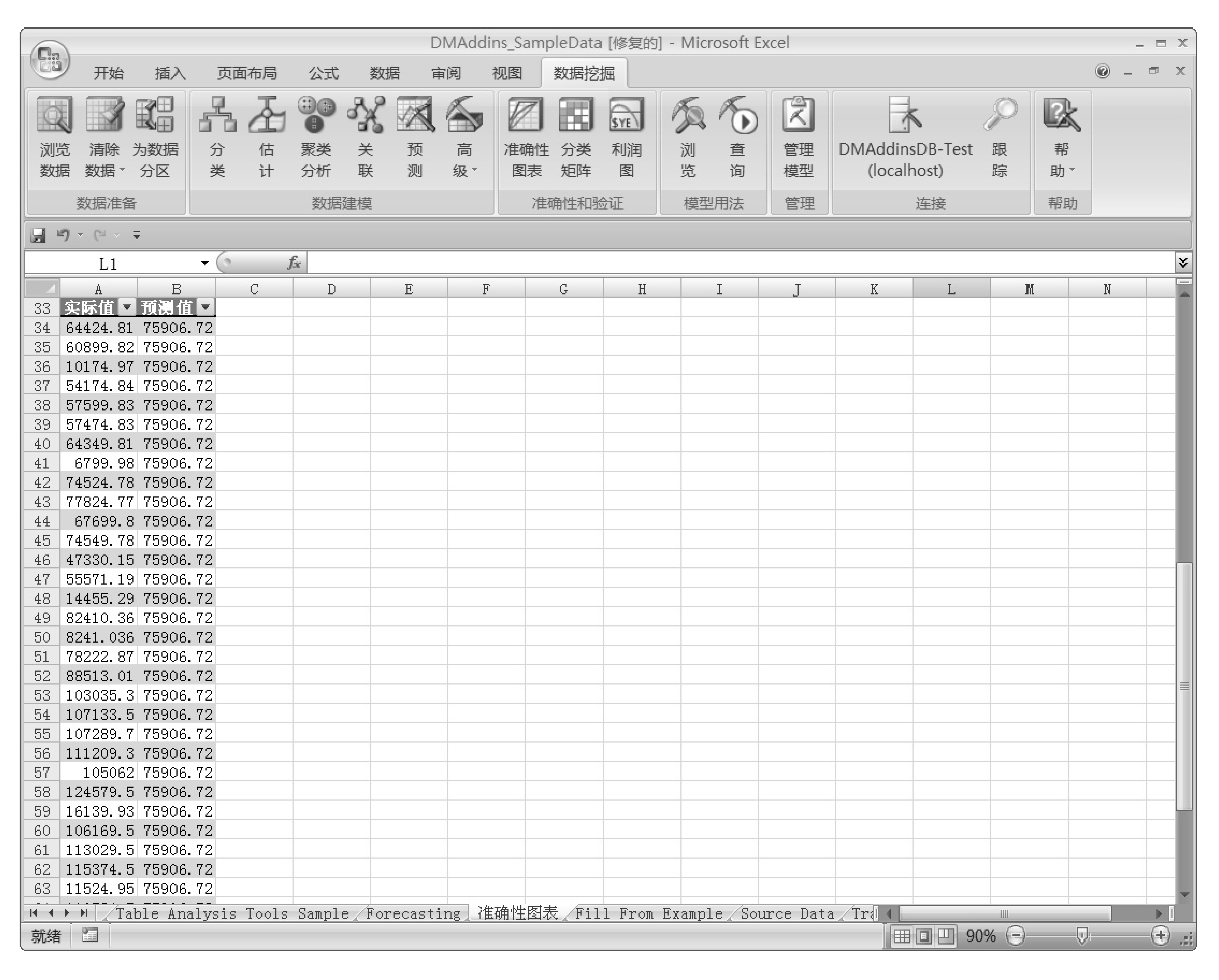
Step16：显示准确性图表，如图 11-16 所示。

Step17：显示预测值，如图 11-17 所示。



图 11-15　【指定关系】窗口

图 11-16　准确性图表



	A	B
33	实际值	预测值
34	64424.81	75906.72
35	60899.82	75906.72
36	10174.97	75906.72
37	54174.84	75906.72
38	57599.83	75906.72
39	57474.83	75906.72
40	64349.81	75906.72
41	6799.98	75906.72
42	74524.78	75906.72
43	77824.77	75906.72
44	67699.8	75906.72
45	74549.78	75906.72
46	47330.15	75906.72
47	55571.19	75906.72
48	14455.29	75906.72
49	82410.36	75906.72
50	8241.036	75906.72
51	78222.87	75906.72
52	88513.01	75906.72
53	103035.3	75906.72
54	107133.5	75906.72
55	107289.7	75906.72
56	111209.3	75906.72
57	105062	75906.72
58	124579.5	75906.72
59	16139.93	75906.72
60	106169.5	75906.72
61	113029.5	75906.72
62	115374.5	75906.72
63	11524.95	75906.72

图 11-17　预测值

第 12 章　线 性 回 归

12.1　基 本 概 念

当某种现象的变化及其分布特性清楚后，需要分析变化发生的原因及其影响因素。在研究变量 X 与 Y 间相互关系时，如果变量 X 可以自由变动，则可用各种试验设计探讨 X 对 Y 的影响；但如果 X 不能自由变动，可用事先求得的 X 与 Y 间的关系来推测 Y 值。

相关分析法和回归分析方法一直都是生物统计学的重要方法。早在 1885 年，高登（F. Galton）在 regression towards mediocrity in hereditary stature 一文中发表根据父母身体特性预测子女身体特性的研究结果。他发现“身高偏高的父母，其子女平均身高要低于他们父母的平均身高；相反地，身高偏矮的父母，其子女平均身高却要高于他们父母的平均身高”。他用 regression 来表示这种效应。因此，将用一个变量去预测另一变量的方法称为“回归分析”。“回归”一词本有其特殊意义，现已将其一般化，用以描述两个或两个以上变量间的关系。所以，回归分析是用以一个或多个自变量来描述、预测或控制某一特定因变量。

对于比较简单的变量间的关系，有时可以凭着过去的经验与直觉来判断；但是对于那些比较复杂或需要精确结果的，就需要依赖客观的统计方法来了解它们之间的关系了。在统计学上用来研究这些关系的统计方法，除了方差分析还有回归分析、相关分析等。

回归分析主要用于了解自变量与因变量间的数量关系。主要目的是了解自变量与因变量关系的方向及强度，并用自变量建立模型对因变量做预测，此外还可以用于分类。

回归分析按照自变量的个数可以分为简单回归分析和多元回归分析。回归分析中变量的选择原则是依相关理论或已有的研究经验和判断。

回归分析步骤：

① 由分布图的情况或专门学科的知识，设定数学模型。

② 用最小平方法推导正规方程。

③ 解出回归方程。

④ 用图示法验证所拟合的回归预测值与观测值的分布是否一致，来确定模型是否合理，如图 12-1 所示。

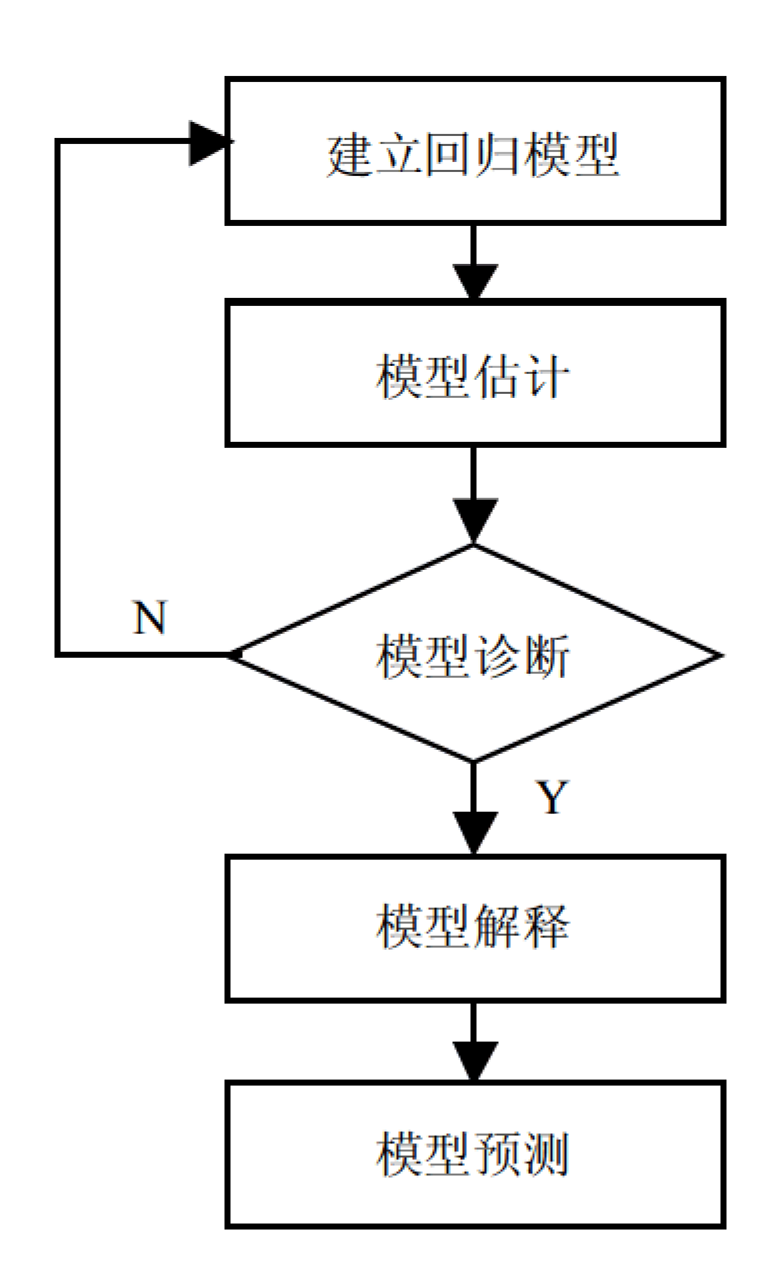


图 12-1　回归分析步骤

12.2 简单回归分析

1．模型假设和估计

假设简单回归模型可用下式表示：

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i, i=1,\cdots,n$$

其中：y_i为因变量，x为自变量，ε_i为误差项，β_0,β_1为回归系数，其中β_0为截距项，β_1为模型的斜率。

误差项代表可能的偏差。回归模型假设的基本思想是误差项来自某一个正态分布$N(0,\sigma^2)$。

回归模型基本假设为：

① 正态分布：对任一固定 x 值，Y 是一个随机变量，有确定的概率分布$Y|X\sim N(u_{y|x},\sigma^2_{y|x})$。

② 独立性：y之间相互独立。

③ $u_{y|x}$是x的线性函数，即$u_{y|x}=\beta_0+\beta_1 x$。

④ 方差齐次性（homoscedasticity）：对于任意的x，有$\sigma^2_{y|x}=\sigma^2$。

简单线性回归分析中最重要的是估计回归系数，估计的方法通常采用普通最小平方法（ordinal least squares method，OLS），也就是使散点图上的所有观测值到回归直线距离的平方和最小。对任一给定的自变量值x_i而言，其相应的估计值表示为$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$。利用最小平方法所得的$\hat{\beta}_0$与$\hat{\beta}_1$值，将使得因变量的观测值$y_i$与因变量的估计值$\hat{y}_i$之间的离差平方和为最小，即$\min\sum(y_i-\hat{y}_i)^2$。

普通最小平方法的推导：

$$\text{SSE}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2$$

分别对$\hat{\beta}_0$、$\hat{\beta}_1$微分，并令其为 0：

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}_0}=0，\frac{\partial \text{SSE}}{\partial \hat{\beta}_1}=0$$

$$\Rightarrow\begin{cases}\sum y=n\hat{\beta}_0+\hat{\beta}_1\sum x\\ \sum xy=\hat{\beta}_0\sum x+\hat{\beta}_1\sum x^2\end{cases}\quad\text{（正规方程）}$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}$$

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_1 \overline{x}$$

最小平方法可提供描述自变量与因变量关系的最佳近似直线。由最小平方法建立的直线方程称为估计回归线或估计回归方程，并以 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 表示，$\hat{y}_i$ 是 y 的预测值或估计值。两者之差反映了估计的误差，第 i 个观察值之差为 $e_i = y_i - \hat{y}_i$，此差值称为第 i 个观察值的残差（residual）。

（1）对 σ^2 的估计

σ^2 是误差项 ε 的方差，通常以误差平方和 SSE 求得 σ^2 的估计值。以 $\hat{\sigma}^2$ 估计 σ^2：

$$\hat{\sigma}^2 = S_{y|x}^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i) = \frac{n-1}{n-2}(S_y^2 - \hat{\beta}_1 S_x^2)$$

其中：

$$S_y^2 = \frac{\sum y_i^2 - n\left(\overline{y}\right)^2}{n-1}, \quad S_x^2 = \frac{\sum x_i^2 - n\left(\overline{x}\right)^2}{n-1}$$

（2）对 β_1 的统计推断：$H_0 : \beta_1 = \beta_1^*, H_1 : \beta_1 \neq \beta_1^*$

在线性回归模型中有：

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (x_i - \overline{x})^2}\right)$$

检验统计量为：

$$t = \frac{\beta_1 - \beta_1^*}{\frac{S_{y|x}}{S_x\sqrt{n-1}}} \sim t_{n-1,1-\alpha/2}$$

如果 $|Z| > t_{n-2,1-\alpha/2}$，则拒绝 H_0。

其中 β_1 的 100×(1−a)%置信区间上下界为 $\beta_1 \pm t_{n-2,1-\alpha/2}\frac{S_{y|x}}{S_x\sqrt{n-1}}$。

（3）对 β_0 的统计推断：$H_0 : \beta_0 = \beta_0^*, H_1 : \beta_0 \neq \beta_0^*$

$$\beta_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum (x_i - \overline{x})^2}\right)\right)$$

检验统计量为：

$$t=\frac{\hat{\beta}_0-\beta_0^*}{S_{y|x}\sqrt{\dfrac{1}{n}+\dfrac{\overline{x}^2}{(n-1)S_x^2}}}\sim t_{n-2,1-\alpha/2}$$

其中 β_0 的 100×(1−a)%置信区间为：

$$\left[\hat{\beta}_0-t_{n-2,1-\alpha/2}S_{y|x}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{(n-1)S_x^2}},\hat{\beta}_0+t_{n-2,1-\alpha/2}S_{y|x}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{(n-1)S_x^2}}\right]$$

（4）回归系数的意义

回归系数表示当自变量 X 发生一个单位的变化时，因变量 Y 相应发生的平均变化量。假设变量 Y=“销售量”和变量 X=“广告投资”的回归方程为 $\hat{Y}=120+0.24X$。其意思是：平均来说，如果“广告投资”X 增加 100 万元，则“销售量”Y 将增加约 24 万元。$\hat{\beta}_0$=120 表示当广告投资 $X=0$ 时，平均的销售量；$\hat{\beta}_1$=0.24 表示等于 X 增加一个单位（1 万元）时 Y 平均的增加量。

2．回归模型拟合优度检验

首先介绍用来衡量估计回归方程拟合优度（goodness of fit）的判定系数（coefficient of determination）。用普通最小平方法可求出使因变量的观测值 y_i 与其预测值 $\hat{y}_i$ 之间的离差平方和最小的 $\hat{\beta}_0$ 与 $\hat{\beta}_1$。因此普通最小平方法所处理的平方和，常被称为误差平方和或残差平方和，以 SSE 表示，是由未知原因所引起的变异。

$$\text{误差平方和：}\ \mathrm{SSE}=\sum(y_i-\hat{y}_i)^2$$

与平均数有关的平方和（记为 SST），也就是总方差，定义如下：

$$\text{总平方和：}\ \mathrm{SST}=\sum(y_i-\overline{y})^2$$

为衡量估计回归直线的预测值 $\hat{y}$ 与 $\overline{y}$ 的差异，需要计算回归平方和（sum of squares due to regression，记做 SSR），它表示由自变量 X 回归引起的方差，即由回归方程解释的方差。回归平方和定义如下：

$$\text{回归平方和：}\ \mathrm{SSR}=\sum(\hat{y}_i-\overline{y})^2$$

SSE、SST 与 SSR 的关系为 SST = SSR+ SSE。

接下来探讨 SSE、SST 与 SSR 如何测量回归关系的拟合优度。如果各观测值均落在最小平方线上，这是最佳拟合的情况，直线通过每一点，所以 SSE=0。因此，在完全拟合情况下，SSR 与 SST 必然相等，即 SSR/SST=1。从另一方面来看，拟合优度不好则导致较大的 SSE。然而，由于 SST+SSR=SSE，所以当 SSR=0 时，SSE 为最大（拟合优度最差）。在这种情况下，估计回归方程无法预测 y。因此，拟合效果最差的回归模型将使 SSR/SST=0。

用 SSR/SST 评估回归关系的拟合优度，判定系数介于 0～1 之间，记做 r^2。其值越接近 1，表示拟合优度越好。

$$\text{判定系数：}\ r^2=\frac{\mathrm{SSR}}{\mathrm{SST}}=1-\frac{\mathrm{SSE}}{\mathrm{SST}}$$

SSR 为可由回归方程解释的 SST 部分。可将判定系数理解为回归模型对 SST 的解释程度。当以百分比表示时，判定系数可解释为在 SST 中，回归方程可以解释的百分比，即自变量 X 解释了因变量变动的百分数。较大的 r^2 值仅表示该回归模型提供较好的拟合，但不能仅依 r^2 的大小来判断 X 与 Y 之间的关系是否为统计显著。若要下这类结论，必须考虑样本大小与最小平方估计的渐进抽样分布的性质。

对社会科学数据而言，即使 r^2 低如 0.25，通常可视为有用的。对自然医疗科学数据而言，经常发现高于 0.60 的 r^2 值。事实上，有时更能见到 r^2 值高于 0.90 的情形。

12.3　多元回归分析

多元回归是简单线性回归的推广，模型包含一个因变量和 K（K≥2）个自变量。例如，在研究“销售量 Y”的变化时，只考虑“广告投资 X_1”可能不够，可能还要考虑“销售人员的数量 X_2”、“特定产品的价格 X_3”、“个人可支配所得 X_4”等其他变量，此时采用多元回归分析是比较妥当的。需要注意的是，如果因变量是定性变量，例如因变量“购买意向 Y”为二分变量时，也就是 $Y=1$ 表示肯定购买，$Y=0$ 表示不一定购买，则要采取 Logistic 回归分析。

多元回归分析可以达到以下目的：

① 了解因变量和自变量之间的关系是否存在，以及这种关系的强度。也就是以自变量所解释的因变量的变异部分是否显著，且因变量变异中有多大部分可以由自变量来解释。

② 估计回归方程，求在自变量已知的情况下因变量的理论值或预测值，达到预测目的。

③ 评价特定自变量对因变量的贡献，也就是在控制其他自变量不变的情况下，该自变量的变化所导致的因变量变化情况。

④ 比较各自变量在拟合的回归方程中相对作用的大小，寻找最重要的和比较重要的自变量。

多元回归模型，其公式如下：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\beta_kX_k+\varepsilon$$

该模型可以用下面的回归方程来估计：

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\hat{\beta}_3X_3+\cdots+\hat{\beta}_kX_k$$

其中 β_0 代表截距，β_1 代表回归系数（也就是偏回归系数），一般都是通过常用的统计软件来估计，统计软件还将给出标准回归系数和对应的标准误差，这些统计量与简单回归中给出的相应的统计量的意义是一致的。

1. 回归效果的评估

对所有自变量与因变量之间的直线回归关系的拟合程度，可以用类似于简单回归中决定系数的统计量 R^2 来度量，其公式如下：

$$R^2=\frac{\text{SSR}}{\text{SST}}=1-\frac{\text{SSE}}{\text{SST}}$$

SST（Y 的总变异）= SSR（可由回归方程解释的变异）+ SSE（不可解释的变异）

其中：

$$\text{SST}=\sum(Y-\overline{Y})^2,\quad \text{SSR}=\sum(\hat{Y}-\overline{Y})^2,\quad \text{SSE}=\sum(Y-\hat{Y})^2$$

称 R^2 为决定系数或复相关系数 R 的平方。R 和 R^2 具有以下的意义和性质：

① R 也可以看成是实际值 Y 和预测值 $\hat{Y}$ 之间的简单相关系数 r。

② 决定系数 R^2 不会小于因变量 Y 和任一个自变量 X 之间的最大的决定系数 r^2，即 $R^2 \geqslant \max\left\{r_1^2, r_2^2, \cdots, r_k^2\right\}$，其中 r_i^2 为 Y 与 X_i 的判定系数。

③ 自变量 $X_1, X_2, \cdots, X_k$ 之间相互相关程度越低，R^2 的值就可能越高。

④ 如果自变量 $X_1, X_2, \cdots, X_k$ 之间是统计上独立的，则 R^2 就等于所有自变量与因变量的决定系数之和，即 $R^2=\sum r_i^2$。

当回归方程中自变量的个数持续增加时，R^2 值不会减小；不过，在前几个自变量之后，再增加自变量也不会对 R^2 有多大的贡献。因此，不难发现当 R^2 很大时，应考虑是否是增加变量导致的。为避免此问题产生，应加以调整，即按照自变量的个数和样本量对 R^2 进行如下的调整：

$$R_{\text{adj}}^2=1-\frac{\text{SSE}/(n-k)}{\text{SST}/(n-1)}=R^2-\frac{k(1-R^2)}{(n-k-1)}$$

此时称 R_{adj}^2 为调整决定系数（adjusted coefficient of determination）。

2. 回归模型的假设检验

回归模型的显著性检验包括：① 对整个回归方程的显著性检验；② 对回归系数的显著性检验。

对整个回归方程的显著性检验的假设为“总体的决定系数 ρ^2 为零”，这个零假设等价于“所有的总体回归系数都为零”，即：

$$H_0:\ \rho^2=0 \quad 或 \quad H_0:\ \beta_1=\beta_2=\cdots=\beta_k=0$$

检验的统计量为 R^2，最终检验统计量为 F 比值，计算公式为：

$$F=\frac{\text{SSR}/k}{\text{SSE}/(n-k-1)}=\frac{R^2/k}{(1-R^2)/(n-k-1)}$$

自由度=（$k, n-k-1$）

F 比值的意义实际上是“由回归解释的方差”与“不能解释的方差”之比，由总变异的分解式可以看到回归方差的显著性检验与方差分析的概念是类似的。因此也称上述检验过程为应用于回归的方差分析，如表 12-1 所示。

表 12-1 多元回归的 ANOVA

变异的来源	变　异	自　由　度	方　差	F 比值
可以解释　（回归）	$\sum(\hat{Y}-\bar{Y})^2$	k	$MSR=\dfrac{SSR}{k}$	$F=\dfrac{MSR}{MSE}$
不可解释　（残差）	$\sum(Y-\hat{Y})^2$	$n-k-1$	$MSE=\dfrac{SSE}{(n-k-1)}$	
总计	$\sum(Y-\bar{Y})^2$	$n-1$		

对某个回归系数 β_i 的显著性检验的零假设为：

$$H_0:\beta_i=0$$

检验的最终统计量仍为 T 统计量：

$$t_i=\frac{\beta_i}{\mathrm{SE}(\beta_i)},i=1,\cdots,k$$

3．回归变量的选择

在建立回归方程时，可能会涉及很多自变量。然而有些变量可能并不重要，太多的变量会使模型变得过于复杂。因此，需要对大量的自变量进行必要的筛选，用尽可能少的自变量去解释因变量中最大比例的变异。选择回归变量的常用方法主要有以下几种。

① 所有可能回归法（all possible regression procedure）：将所有可能的自变量全部加入，进行回归分析。

② 向前选择法（forward selection）：将自变量逐个加入回归模型，检验其是否满足某个事先规定的标准；如果满足该标准，则将此变量加入回归模型，否则就不保留。例如，根据待加入变量对可解释的方差贡献的大小，可以规定“重要的”变量加入方程所需的最小 F 比值（如 F=3.84）或最大概率值 P（如 P=0.05）。

③ 向后淘汰法（backward elimination）：先将全部自变量都加入回归模型中，然后逐个检验其是否满足某个事先规定的剔除比值。如果满足该标准，则将此变量从回归模型中剔除，否则就保留。例如，根据变量对可解释的方差贡献的大小，可以规定将“不重要的”变量从方程中剔除的 F 比值的上限（如 F=2.71）或概率值 P 的下限（如 P=0.10）。

④ 逐步回归法（stepwise regression）：是前两种方法的结合，即根据某些事先规定的标准，逐个加入“重要的”变量，又随时剔除“不重要的”变量，直至既无不显著变量从回归方程中剔除，又无显著变量加入回归方程为止。

注意，按照上述方法得到的回归方程的决定系数 R^2 不一定是最大的，即回归效果不一定是最佳的。由于自变量之间可能相关（即共线性），因此重要的变量有可能被剔除，不重要的变量也有可能被加入。因此，在变量选择的问题上要持慎重的态度，要结合相关的专业知识，考虑各种可能，必要时还可将某些虽然已被剔除，但却“不可缺少的”变量强行加入方程。

12.4　Excel 2007 线性回归

Microsoft 线性回归算法是 Microsoft 决策树算法的特例，即禁止数据分割的决策树（整个回归公式是在单一根节点中建立），但该算法支持连续属性的预测。

Step1：单击【高级】中的【创建挖掘模型】按钮，如图 12-2 所示。

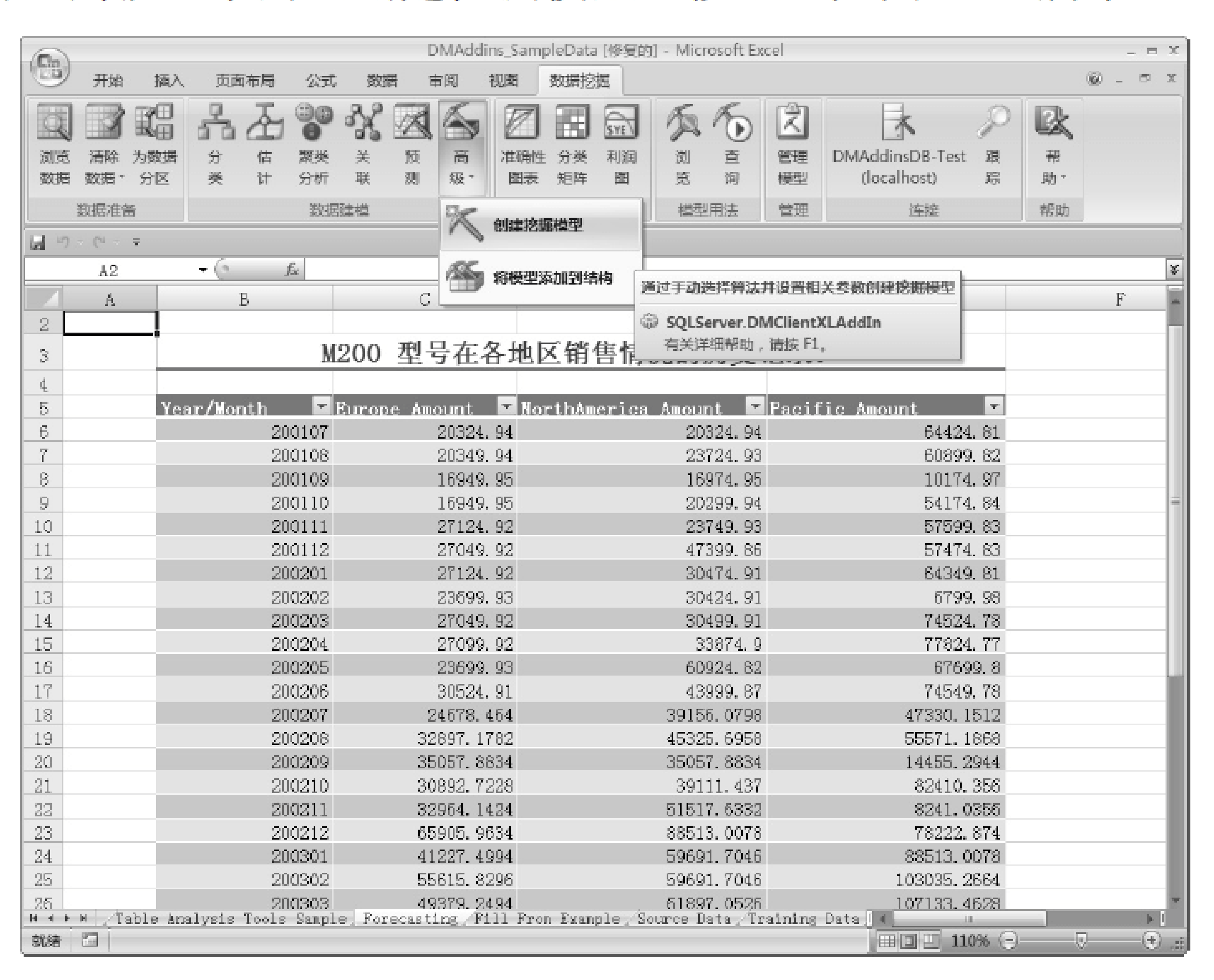


图 12-2　创建挖掘模型

Step2：弹出如图 12-3 所示的【创建模型向导入门】窗口，单击【下一步】按钮。

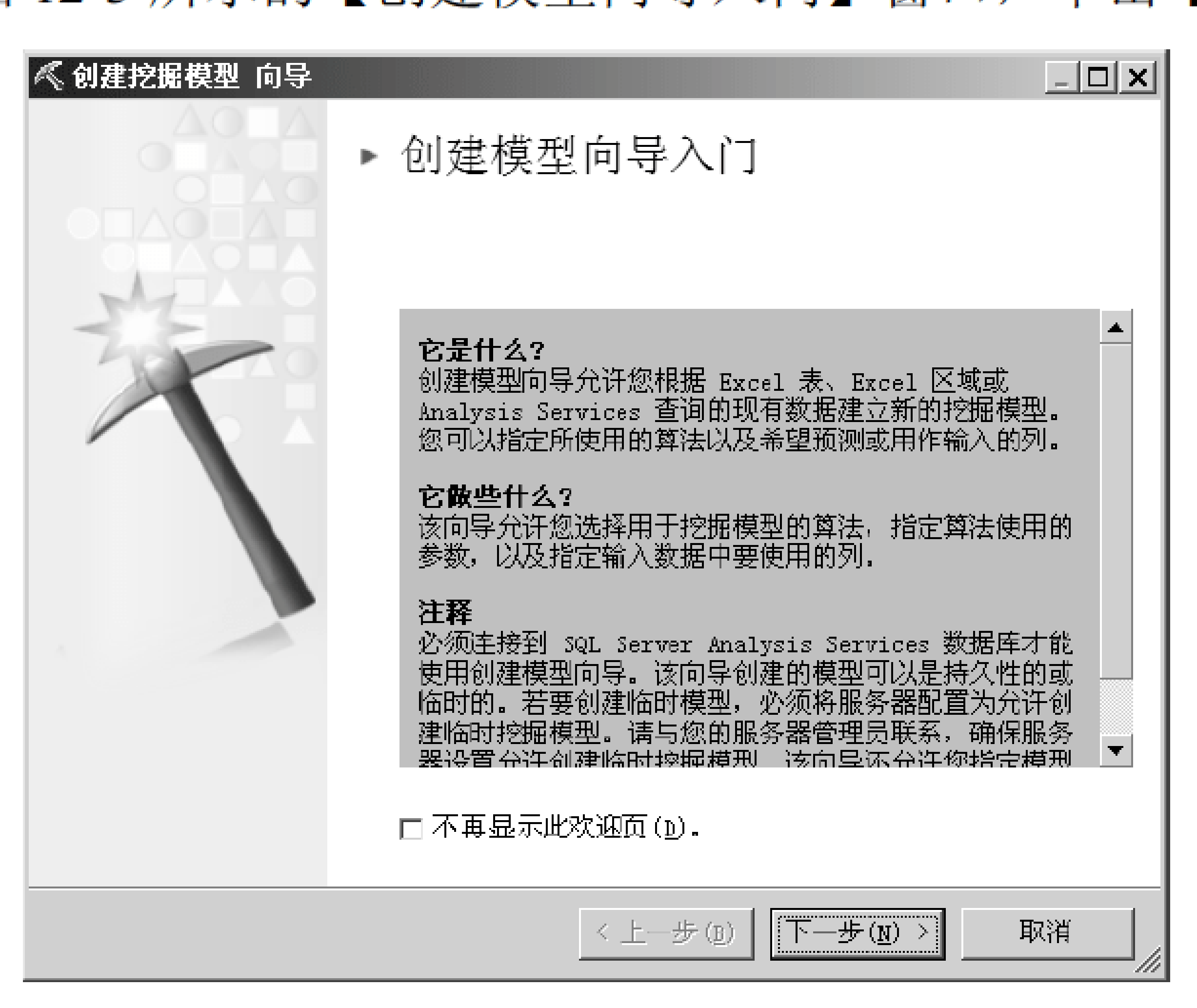


图 12-3　【创建模型向导入门】窗口

Step3：这里选择另外一种不同的数据源——基于 SQL 2005 的 Analysis Services 中的

数据源。选中【Analysis Services 数据源】单选按钮，单击【下一步】按钮，如图 12-4 所示。



图 12-4 选择源数据

Step4：可在【服务器数据源】下拉列表框中选择数据库中已存在的数据，单击红色部分可新增数据来源，如图 12-5 所示。



图 12-5 选择服务器数据源

Step5：在【服务器名称】文本框中输入“localhost”，在【目录名称】下拉列表框中选择 DM_demo，如图 12-6 所示。测试连接无误后单击【确定】按钮。



图 12-6 新建 Analysis services 数据源

Step6：将所需的数据表移到【查询中的列】框内，如图 12-7 所示。



图 12-7 编辑数据源

Step7：在如图 12-8 所示的【选择源数据】窗口中，单击【下一步】按钮。



图 12-8 【选择源数据】窗口

Step8：在如图 12-9 所示的【选择挖掘算法】窗口中，单击【下一步】按钮。

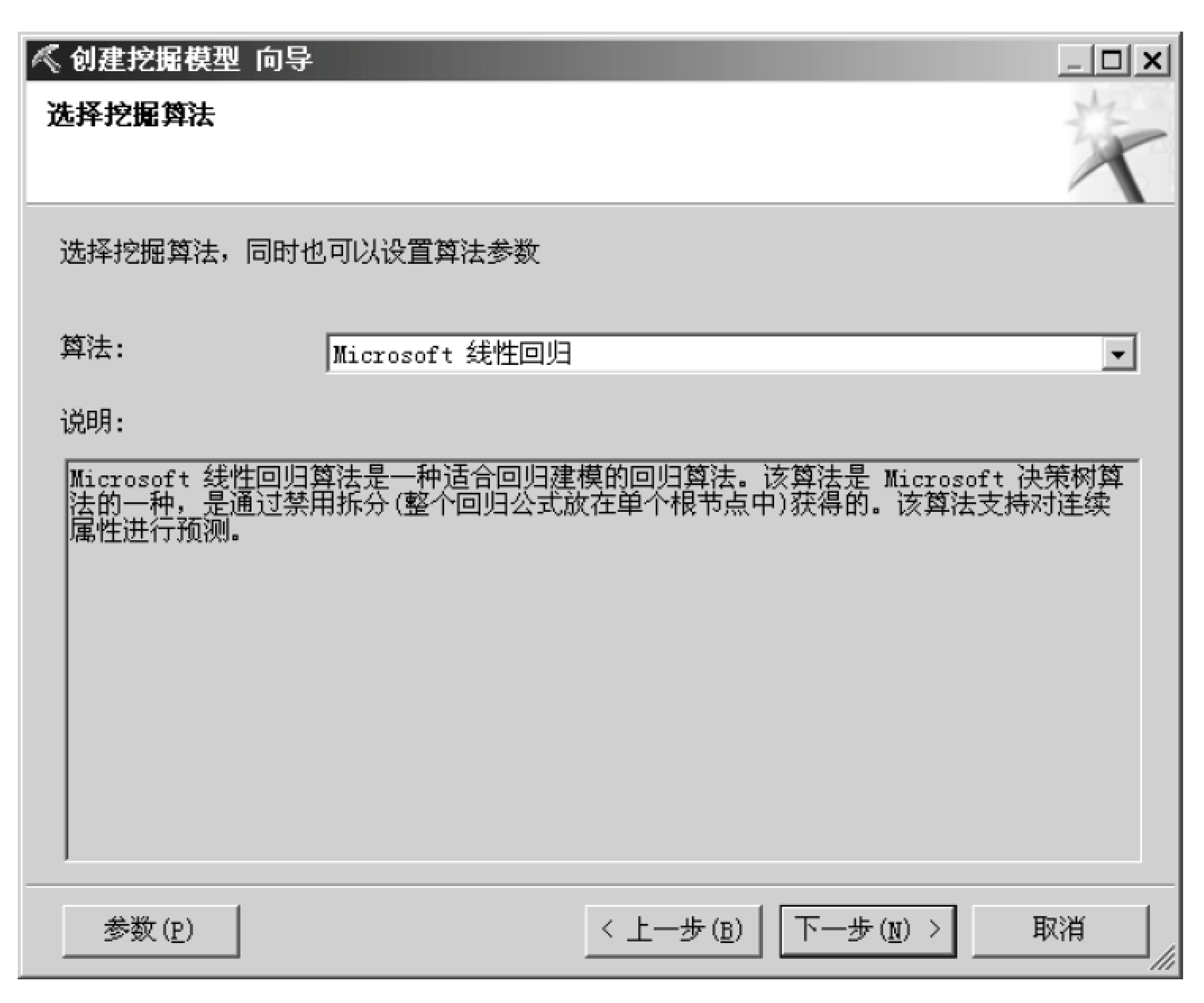


图 12-9 【选择挖掘算法】窗口

Step9：将自变量设定为“输入”，预测变量设定为“仅预测”，把数据中的序号设定为 key，而不使用的变量则设为“不使用”，完成后单击【下一步】按钮，如图 12-10 所示。

Step10：选中【浏览模型】复选框，选中【启用钻取】复选框，单击【完成】按钮，如图 12-11 所示。当然也可以更改结构名称及模型名称。



图 12-10 【选择列】窗口



图 12-11 【完成】窗口

Step11：选择【依赖关系网络】选项卡，若结果有数个变量与预测变量存在关系，则可调整【所有链接】滑块，看出其中关联的强弱程度，如图 12-12 所示。

Step12：单击【数据挖掘】中的【准确性图表】按钮，在如图 12-13 所示的【准确性图表向导入门】窗口中单击【下一步】按钮。

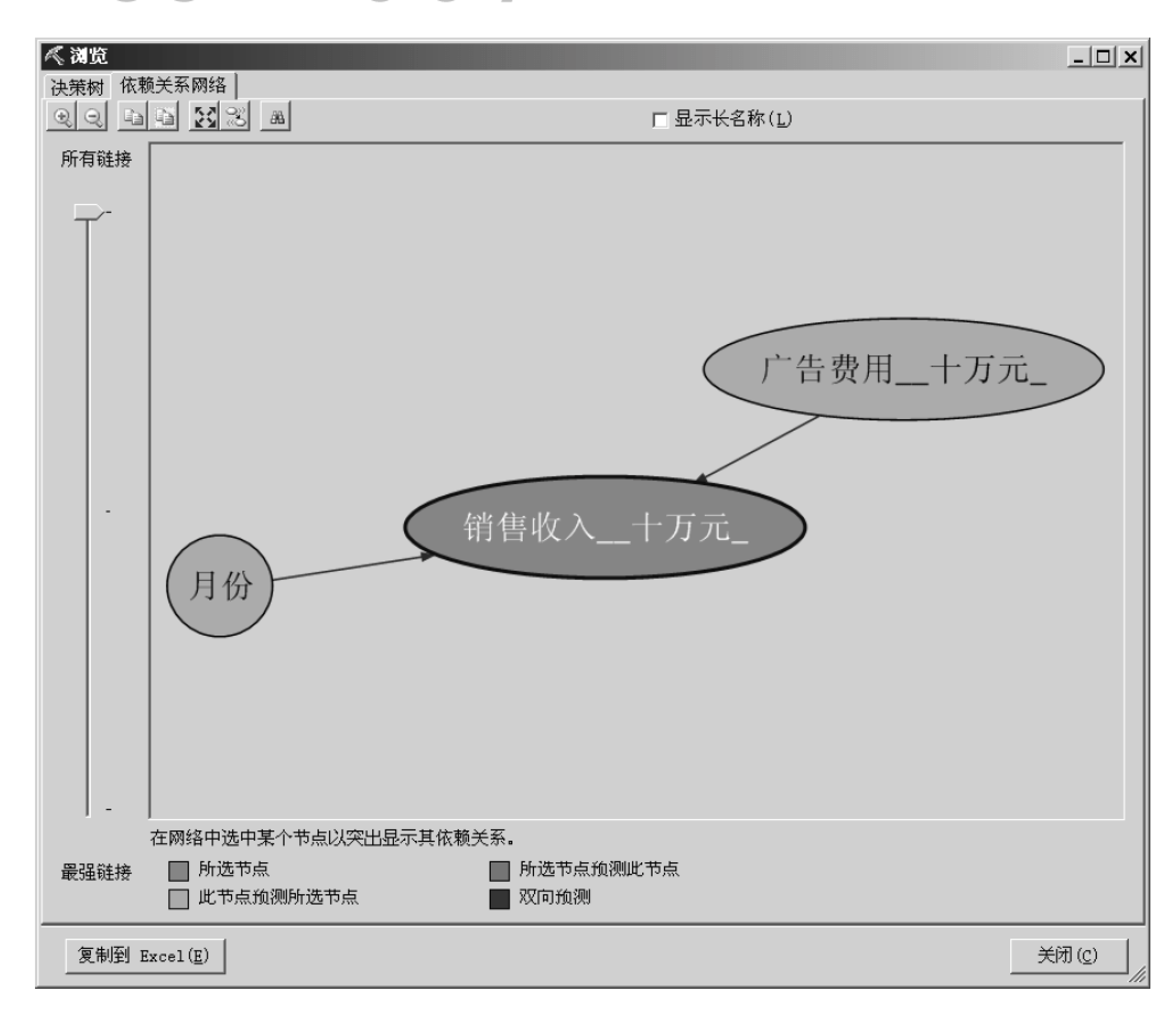


图 12-12 【依赖关系网络】选项卡

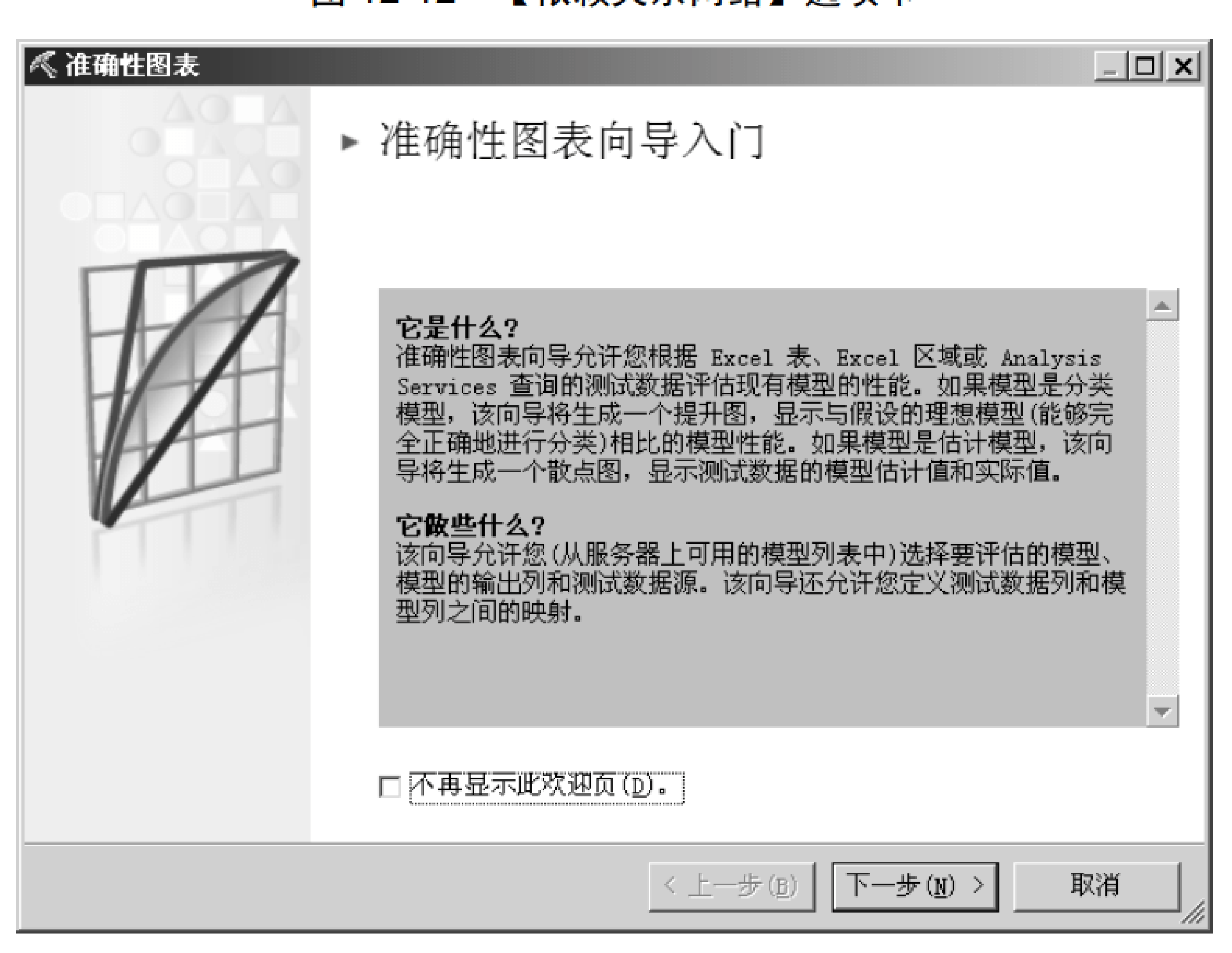


图 12-13 【准确性图表向导入门】窗口

Step13：在模型列表中选择“reg-线性回归_2”，单击【下一步】按钮，如图 12-14

所示。

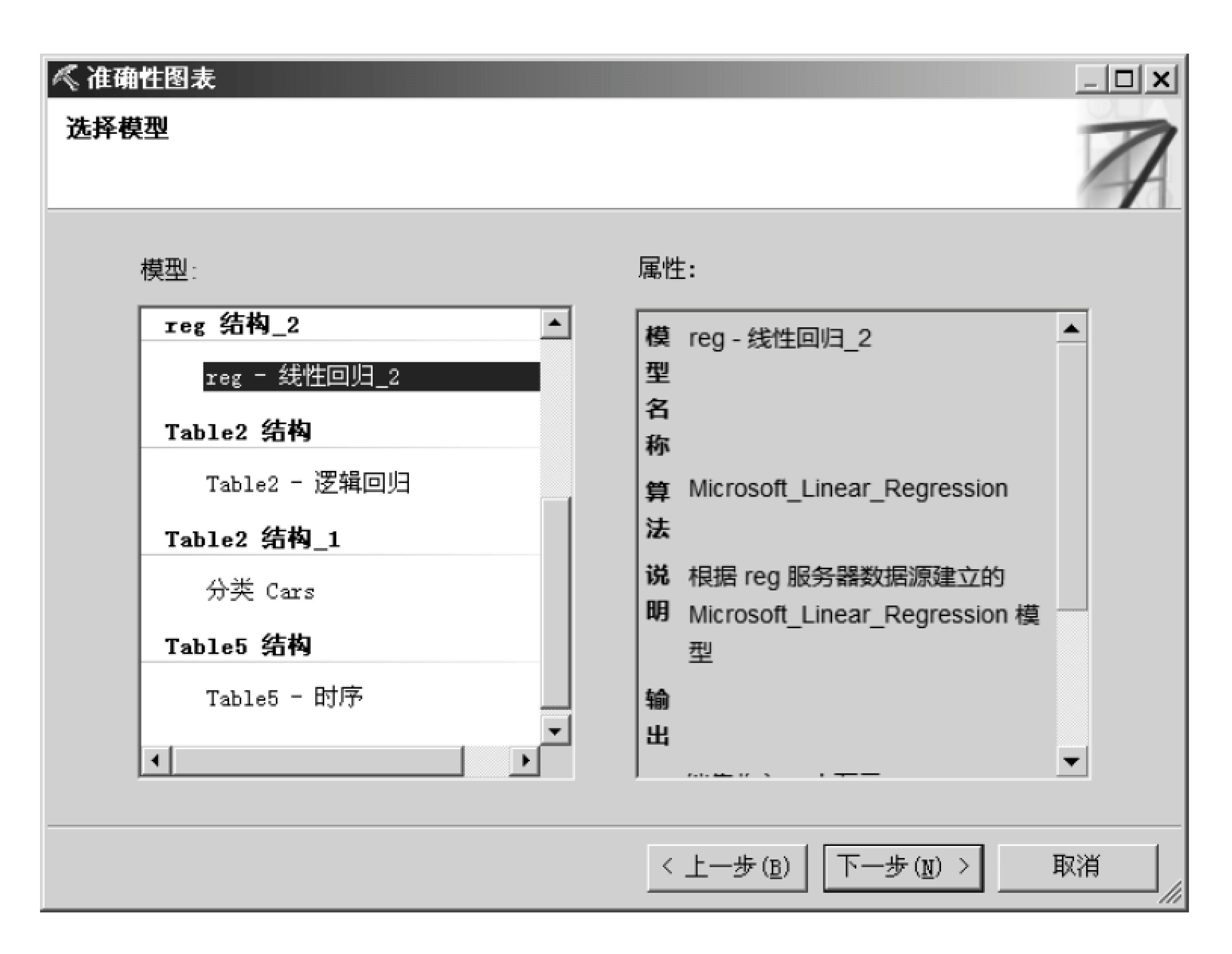


图 12-14 【选择模型】窗口

Step14：在如图 12-15 所示的【指定要预测的列和要预测的值】窗口中，单击【下一步】按钮。

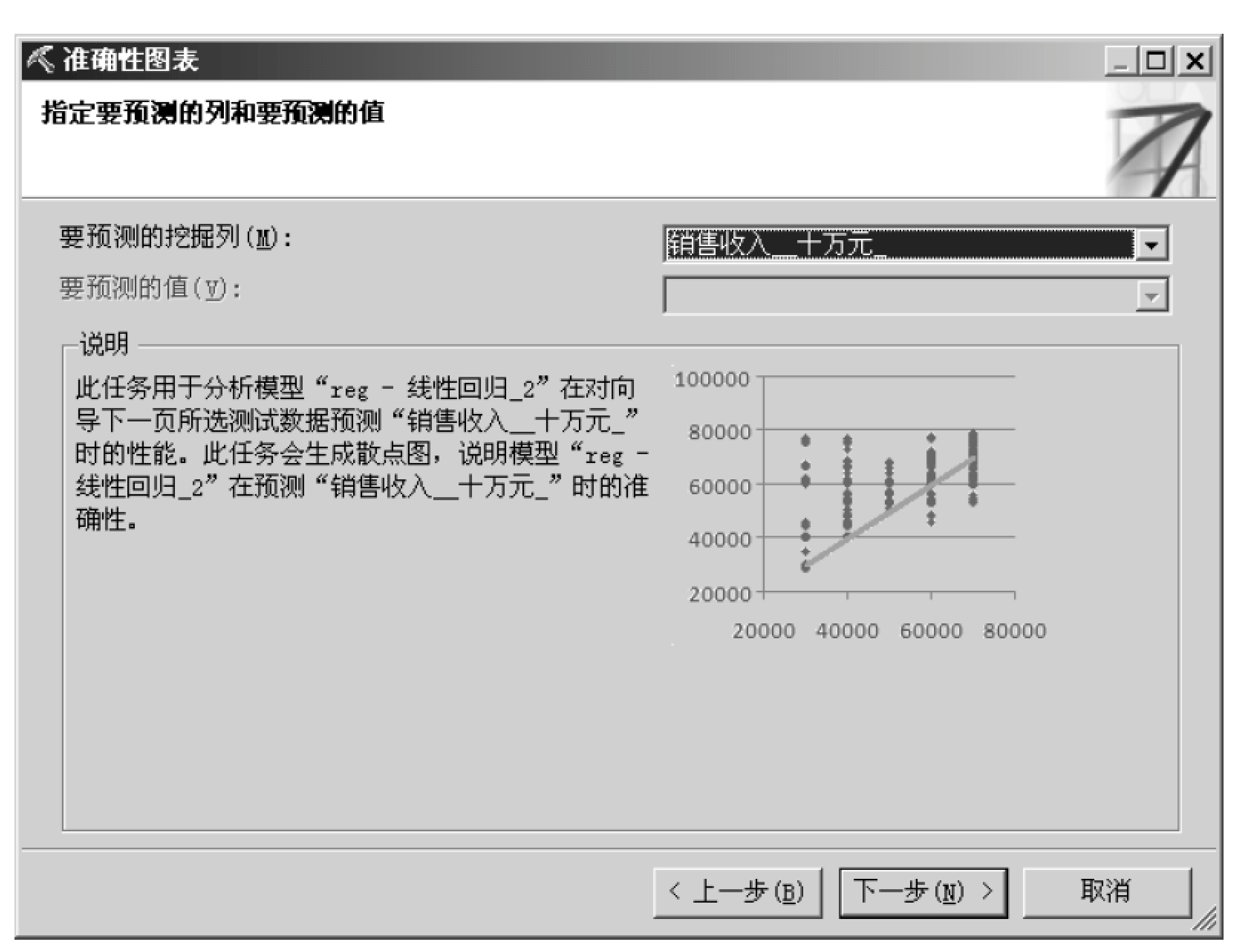


图 12-15 【指定要预测的列和要预测的值】窗口

Step15：选中【Analysis Services 数据源】单选按钮，单击【下一步】按钮，如图 12-16 所示。



图 12-16　选择源数据

Step16：在如图 12-17 所示的【指定关系】窗口中，单击【完成】按钮。



图 12-17　【指定关系】窗口

Step17：得到此模型的准确性图表，如图 12-18 所示。

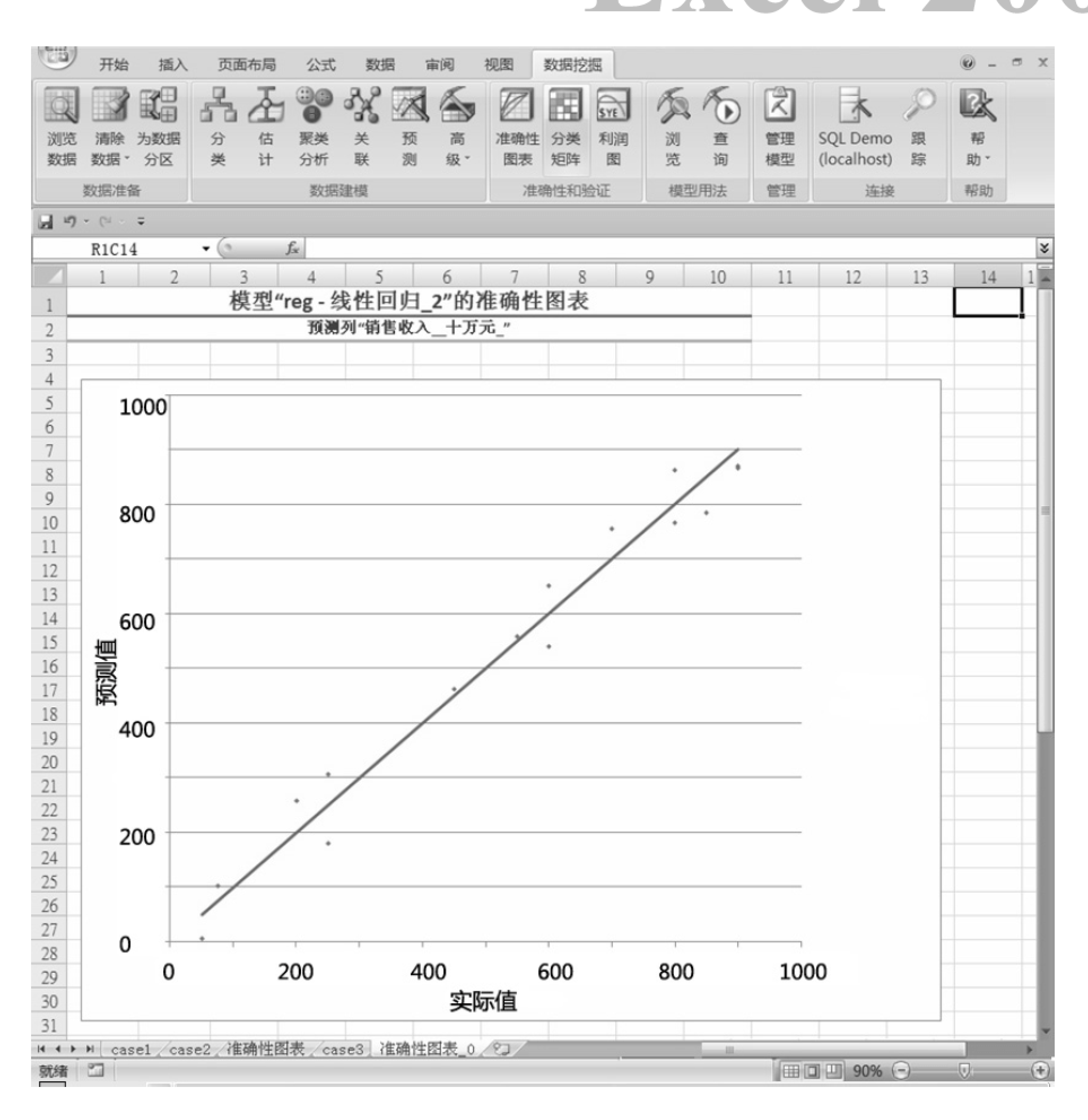


图 12-18　准确性图表

第 13 章　Logistic 回归

13.1　基 本 概 念

Logistic 回归模型用于分析二分类 （binary）或有序（ordinal）的因变量与解释变量间的关系。Logistic 回归模型中，用自变量去预测因变量在给定某个值（如 1 或 0）的概率。因变量通常显示为二分类中某个值或有次序中的最小值。当因变量取很多不同的值时，如等距尺度（interval scale）或比例尺度（ratio scale）的数据类型时，通常使用简单回归模型而不用 Logistic 回归模型。对一个二分类的因变量 Y，Logistic 回归模型的形式如下：

$$\text{Logit}\ P/(1-P) = \alpha + \beta X$$

$P=\text{Prob}(y=Y \mid X)$：代表在给解释变量矩阵下因变量取值的概率，且 y 代表因变量矩阵 Y 中第一个值。

α 代表截距参数矩阵；β 代表斜率参数矩阵；X 代表解释变量矩阵。

Logistic 回归方程即为第 i 组个别事件概率(P_i)的对数(logit)转换，且转换后的 Logistic 回归模型是解释变量矩阵的一条直线方程。而一般化的模型表示法是用因变量的平均数函数 $g=g(u)$ 来表示它与自变量之间的线性关系，g 称为链接函数（link function）。其他常见的链接函数有 probit function 和 log-log function。logit 函数（logit function）有较易解释的优点，同时它也可用于分析将来或过去曾收集到的数据。

对数线性模型是将列联表中每格的概率（或理论频率）取对数后，分解参数获得的；而 Logistic 模型是将概率比取对数后，再进行参数化获得的。为了较好地理解这一方法，先介绍 logit 变换和 Logistic 分布，然后再回到 Logistic 回归分析。

13.2　logit 变换

人们常常要研究某一事件 A 发生的概率 p，p 值的大小与某些因素有关。例如研究有毒药物的剂量大小与被试验的老鼠的死亡率之间的关系，死亡率 p 随着剂量 x 的增大是增长的。但因 p 的值在[0，1]区间内，所以 p 不可能是 x 的线性函数或二次函数，一般的多项式函数也不适合，这就给此类的回归带来困难。另一方面，当 p 接近于 0 或 1 时，一些因素即使有很大变化，p 值的变化也不会显著。如高可靠性系统，可靠度 p 已是 0.998 了，即使再改善条件、工艺和系统的结构，可靠度的增大只能在小数点后三位或四位。又如灾害性天气发生的概率 p 很小，接近于 0，即使能找到一些刻画它发生前兆的信息，也不可能将 p 值提高很多。从数学上看，就是函数 p 对 x 的变化在 p=0 或 1 附近是不敏感的、缓慢的，而且非线性的程度较高，于是要寻求一个 p 的函数 $\theta(p)$，使得它在 p=0 或 p=1 附近时变

化幅度较大，而函数的形式又不是太复杂。

首先，用 $\frac{\mathrm{d}\theta(p)}{\mathrm{d}p}$ 来反映 θ（p）在 p 附近的变化是合理的，同时在 $p=0$ 或 1 时，$\frac{\mathrm{d}\theta(p)}{\mathrm{d}p}$ 应有较大的值，这自然要考虑：

$$\frac{\mathrm{d}\theta(p)}{\mathrm{d}p} \propto \frac{1}{p(1-p)}$$

接着，将上式取成等式，就有：

$$\frac{\mathrm{d}\theta(p)}{\mathrm{d}p} = \frac{1}{p(1-p)} = \frac{1}{p} + \frac{1}{1-p}$$

再求积分后可得：

$$\theta(p) = \ln\frac{p}{1-p}$$

上式相对的变换称为 logit 变换。很明显 θ（p）在 $p=0$ 与 $p=1$ 附近的变化幅度很大，而且当 p 从 0 变到 1 时，θ（p）从 $-\infty$ 变到 ∞，这样就克服了一开始指出的两点困难。如果 p 对 x 不是线性的关系，θ 对 x 就可以是线性的关系了，这给数据处理带来很多方便。从前式，将 p 由 θ 来表示，就得：

$$p = \frac{e^{\theta}}{1+e^{\theta}}$$

如果 θ 是某些自变量 $x_1,\cdots,x_k$ 的线性函数 $\sum_{i=1}^{k} a_i x_i$，则 p 就是 $x_1,\cdots,x_k$ 的函数：

$$p = \frac{e^{\sum_{i=1}^{k} a_i x_i}}{1+e^{\sum_{i=1}^{k} a_i x_i}}$$

很多教材讨论 Logistic 回归时，都是直接从该式开始的。

13.3 Logistic 分布

如果分布函数满足以下形式：

$F(x)=(1+\mathrm{e}^{-(x-\mu)/\sigma})^{-1}, -\infty < x < \infty$（其中 $-\infty<\mu<\infty, \sigma>0$）

则该分布称为 Logistic 分布。另外，$F(x)$也可表示成：

$$F(x) = \frac{1}{2}\left(1+\tanh(\frac{x-\mu}{2\sigma})\right)$$

其密度函数为：

$$f(x) = \frac{1}{\sigma}\mathrm{e}^{\frac{x-\mu}{\sigma}}\left[1+\exp\left(-\frac{x-\mu}{\sigma}\right)\right]^{-2}$$

再将 p 表示成 $F(x)$的形式：

$$p = 1 - F(x) = \mathrm{e}^{-(x-\mu)/\sigma} \big/ (1 + \mathrm{e}^{-(x-\mu)/\sigma})$$

相应地，$\theta = -\dfrac{x-\mu}{\sigma}$。上式说明了 logit 变换与 Logistic 分布的关系。

上式还说明，Logistic 分布仍然是属于位置—尺度参数族，其中 μ 是位置参数，σ 是尺度参数，这样凡是与位置—尺度参数族有关的结果，均对 Logistic 分布有效。当 $\mu = 0, \sigma = 1$ 时，相应的分布称为标准 Logistic 分布，它的分布函数 F_0（x）与分布密度 f_0（x）为：

$$\begin{cases} F_0(x) = (1 + \mathrm{e}^{-x})^{-1} \\ f_0(x) = \mathrm{e}^{-x} \big/ (1 + \mathrm{e}^{-x})^2 \end{cases}, -\infty < x < \infty$$

很明显，如果考虑：

$$G_0(x) = \mathrm{e}^{x} \big/ (1 + \mathrm{e}^{x}), -\infty < x < \infty$$

则 G_0（x）也是一个 Logistic 分布函数，且有如下关系式：

$$G_0(x) = 1 - F_0(-x) = F_0(x)$$

因此有的教材也从 G_0（x）出发，以它作为标准分布。

13.4　列联表的 Logistic 回归模型

现在来讨论如何将 2×2 表转化为一个 Logistic 的回归模型，现以下例为背景进行分析。

假定吸烟人得肺癌的概率是 p_1，不得肺癌的概率就是 $1-p_1$，不吸烟的人得肺癌的概率是 p_2，不得肺癌的概率为 $1-p_2$。于是经过 logit 变换后：

$$\theta_1 = \ln \frac{p_1}{1-p_1}, \qquad \theta_2 = \ln \frac{p_2}{1-p_2}$$

如果记 θ_2 为 θ，则 $\theta_1 = \theta_1 + (\theta_1 - \theta_2) = \theta + \varDelta$。因此患肺癌是否与吸烟有关，就等价于检验 $H_0 : \varDelta = 0$。

考察了 92 个吸烟者，其中 60 个得肺癌，对于不吸烟的 14 个人中有 3 个得肺癌。更一般地，若考察了 n_1 个吸烟者，得肺癌者有 r_1 个；考察 n_2 个不吸烟者，得肺癌者有 r_2 个，因此 p_1 与 p_2 的估计值分别为 $\hat{p}_1 = \dfrac{r_1}{n_1}$，$\hat{p}_2 = \dfrac{r_2}{n_2}$。令：

$$z_i = \ln \frac{r_i}{n_i - r_i}, \qquad i = 1, 2$$

则可以证明，当 n_i 充分大时，有下述等式成立：

$$E(z_i) = \theta_i, \quad \mathrm{Var}(z_i) = \frac{1}{n_i p_i (1 - p_i)}, \quad i = 1, 2$$

如果写成向量的形式，就是：

$$
\begin{cases}
E\begin{bmatrix} z_1 \\ z_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \theta \\ \varDelta \end{bmatrix} \\
\mathrm{Var}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{n_1 p_1(1-p_1)} & 0 \\ 0 & \dfrac{1}{n_2 p_2(1-p_2)} \end{bmatrix}
\end{cases}
$$

如果 z_1，z_2 是正态变量，这就是 2×2 列联表的 Logistic 回归模型。

一般地，当 n_i 充分大时，z_i 服从渐近正态分布，并将这一类问题的回归称为 Logistic 回归。

13.5 Excel 2007 Logistic 回归

Microsoft Logistic 回归算法实际上是 Microsoft 类神经网络算法的特例，即不包含隐藏层的神经网络。该算法同时支持离散属性和连续属性的预测。

Step1：进入 Excel 的范例，如图 13-1 所示。

DMAddins_SampleData [修复的] - Microsoft Excel

用于分析关键影响因素、检测类别、突出显示异常值和应用场景分析的示例数据。

ID	Marital Sta	Gender	Income	Children	Education	Occupation	Home Owner	Cars
12496	Married	Female	40000	1	Bachelors	Skilled Manual	Yes	0
24107	Married	Male	30000	3	Partial College	Clerical	Yes	1
14177	Married	Male	80000	5	Partial College	Professional	No	2
24381	Single	Male	70000	0	Bachelors	Professional	Yes	1
25597	Single	Male	30000	0	Bachelors	Clerical	No	0
13507	Married	Female	10000	2	Partial College	Manual	Yes	0
27974	Single	Male	160000	2	High School	Management	Yes	4
19364	Married	Male	40000	1	Bachelors	Skilled Manual	Yes	0
22155	Married	Male	20000	2	Partial High School	Clerical	Yes	2
19280	Married	Male	20000	2	Partial College	Manual	Yes	1
22173	Married	Female	30000	3	High School	Skilled Manual	No	2
12697	Single	Female	90000	0	Bachelors	Professional	No	4
11434	Married	Male	170000	5	Partial College	Professional	Yes	4
25323	Married	Male	40000	2	Partial College	Clerical	Yes	1
23542	Single	Male	60000	1	Partial College	Skilled Manual	No	1
20870	Single	Female	10000	2	High School	Manual	Yes	1
23316	Single	Male	30000	3	Partial College	Clerical	No	2
12610	Married	Female	30000	1	Bachelors	Clerical	Yes	0
27183	Single	Male	40000	2	Partial College	Clerical	Yes	1
25940	Single	Male	20000	2	Partial High School	Clerical	Yes	2
25598	Married	Female	40000	0	Graduate Degree	Clerical	Yes	0
21564	Single	Female	80000	0	Bachelors	Professional	Yes	4
19193	Single	Male	40000	2	Partial College	Clerical	Yes	0
26412	Married	Female	80000	5	High School	Management	No	3
27184	Single	Male	40000	2	Partial College	Clerical	No	1
12590	Single	Male	30000	1	Bachelors	Clerical	Yes	0
17841	Single	Male	30000	0	Partial College	Clerical	No	1

图 13-1 Excel 范例

Step2：单击【高级】下的【创建挖掘模型】按钮，如图 13-2 所示。

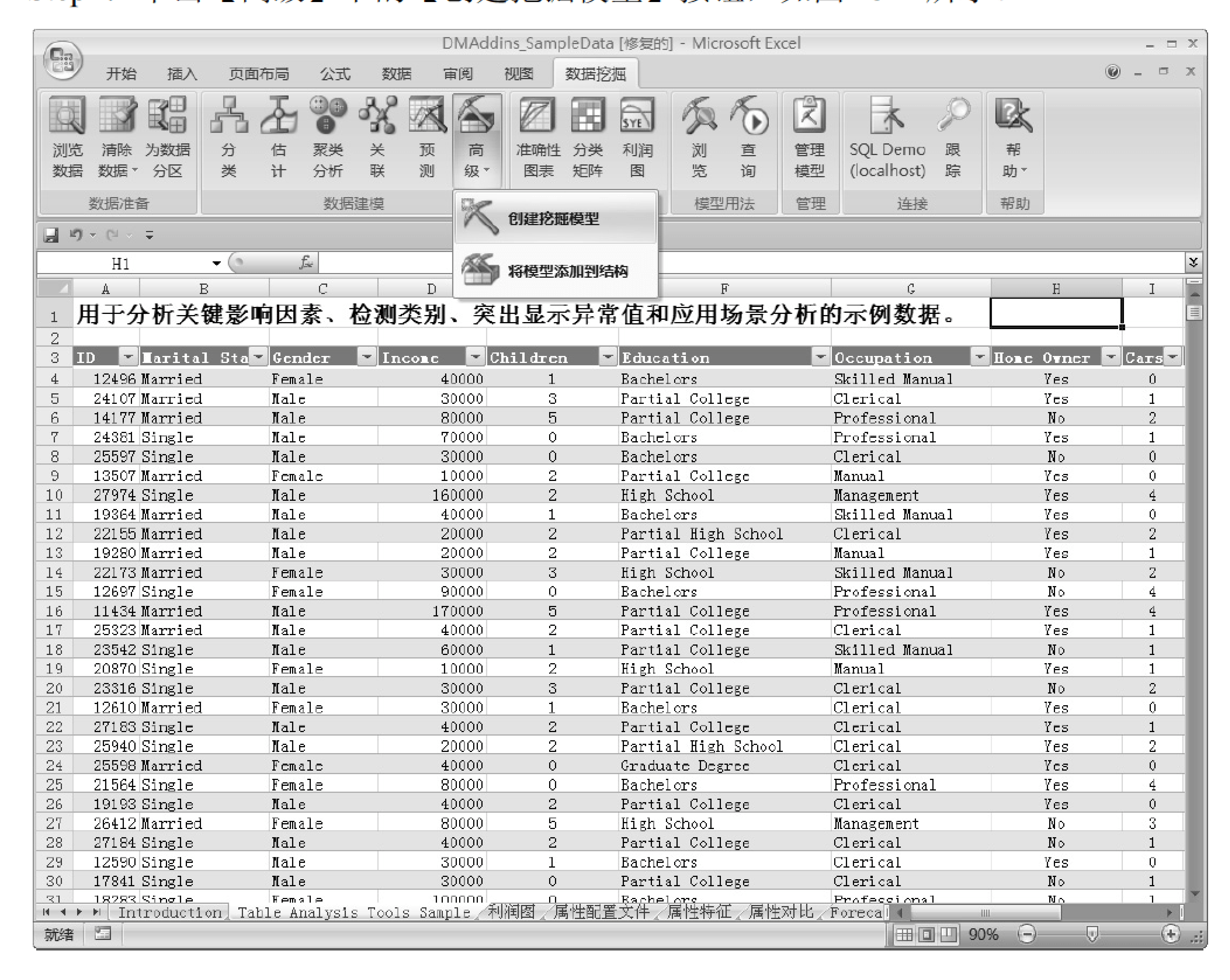


图 13-2　创建挖掘模型

Step3：在如图 13-3 所示的【选择源数据】窗口中，单击【下一步】按钮。



图 13-3　【选择源数据】窗口

Step4：在【数据区域】下拉列表框中选择数据表【'Table Analysis Tools Sample'!'Table2'】，单击【下一步】按钮，如图 13-4 所示。



图 13-4　选择数据表

Step5：在【算法】下拉列表框中选择【Microsoft 逻辑回归】，单击【下一步】按钮，如图 13-5 所示。

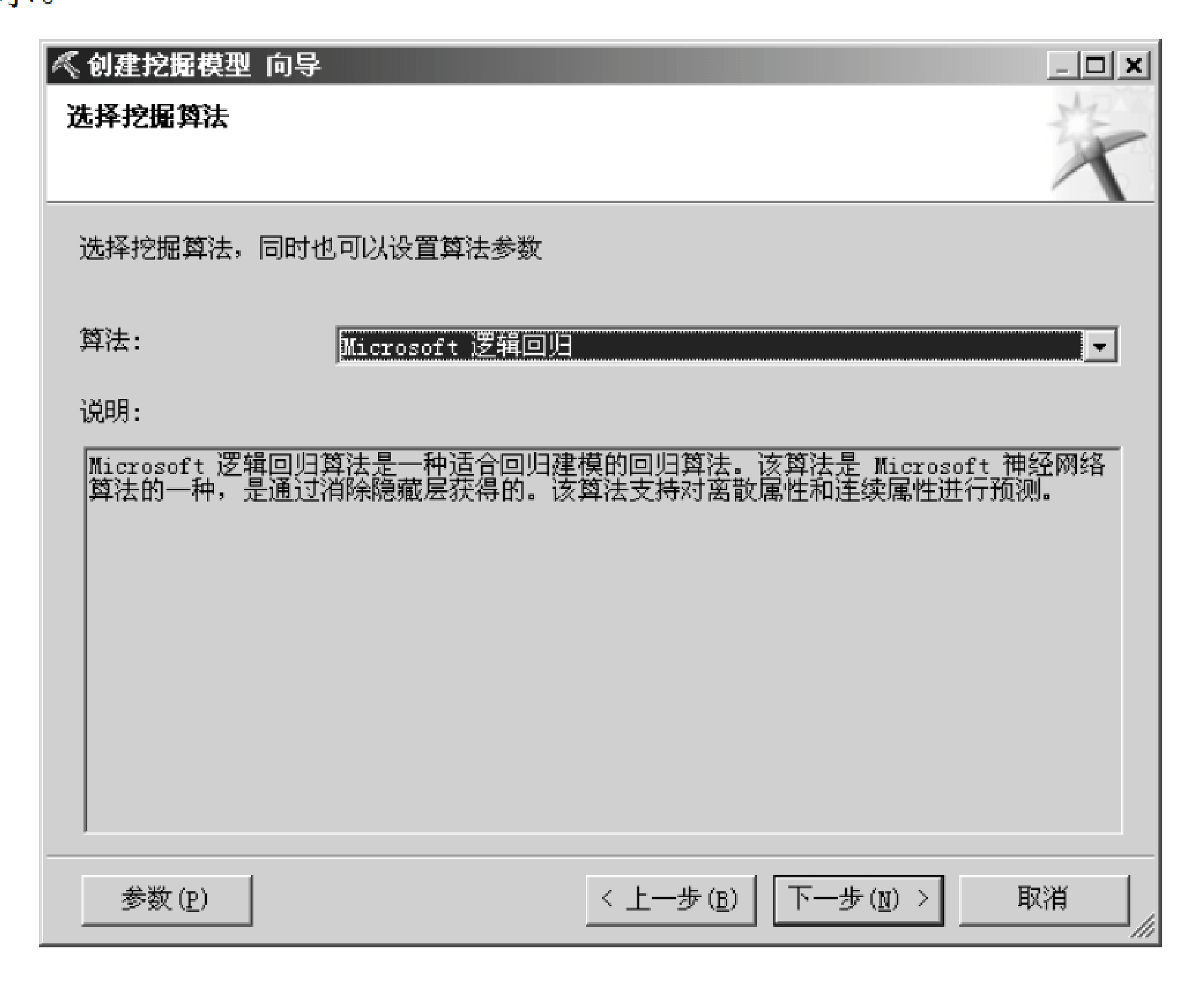


图 13-5　选择挖掘算法

Step6：在如图 13-6 所示的【选择列】窗口中选择被预测变量 Purchased Bike，单击【下

一步】按钮。



图 13-6 【选择列】窗口

Step7：在如图 13-7 所示的【完成】窗口中，选中【浏览模型】复选框，单击【完成】按钮。



图 13-7 【完成】窗口

Step8：在如图 13-8 所示的【浏览】窗口中，可看出在被预测变量 Purchased Bike 中各

变量属性的特征。

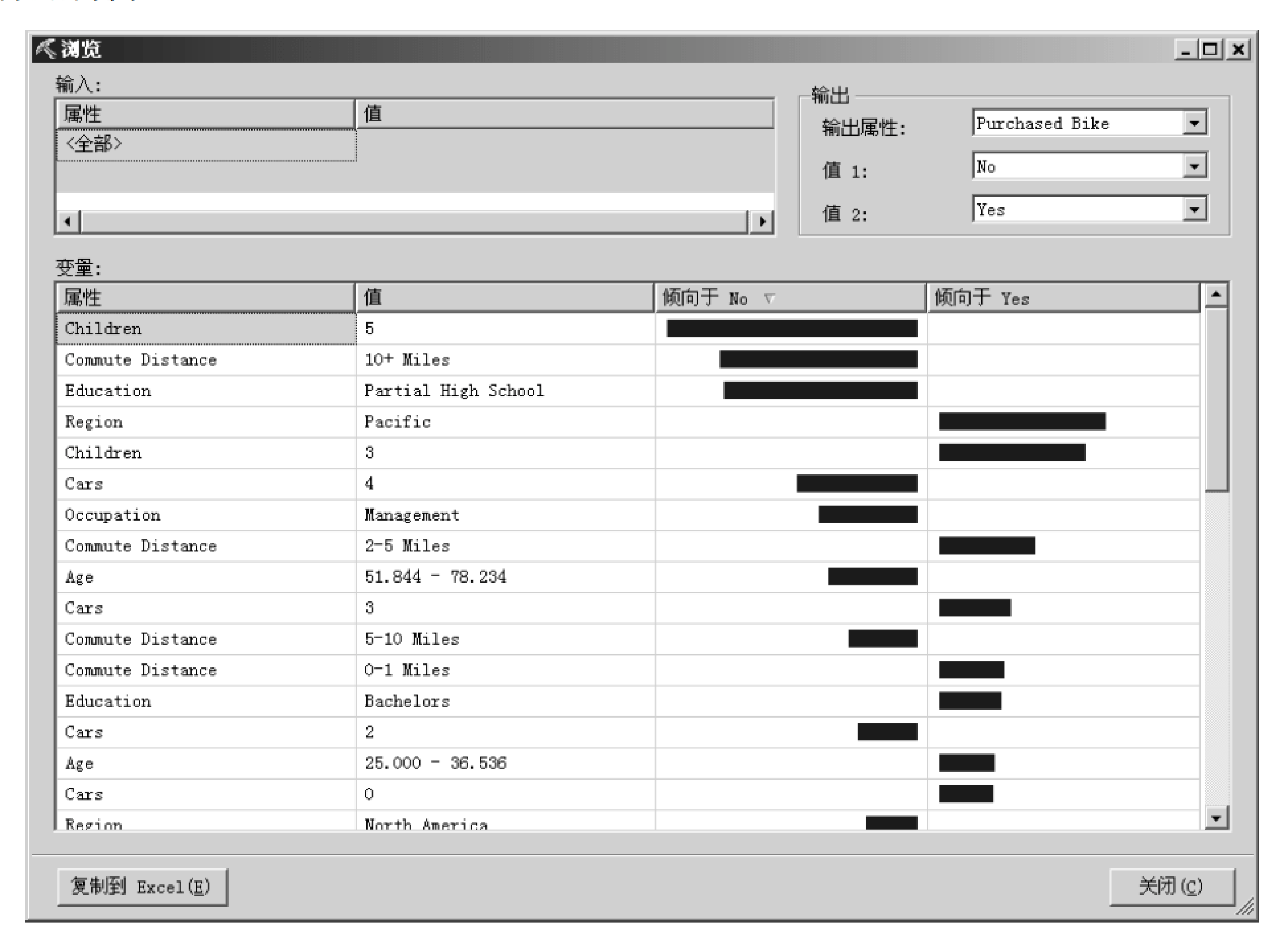


图 13-8 【浏览】窗口

Step9：将列出的 Logistic 回归的特征复制到 Excel，如图 13-9 所示。

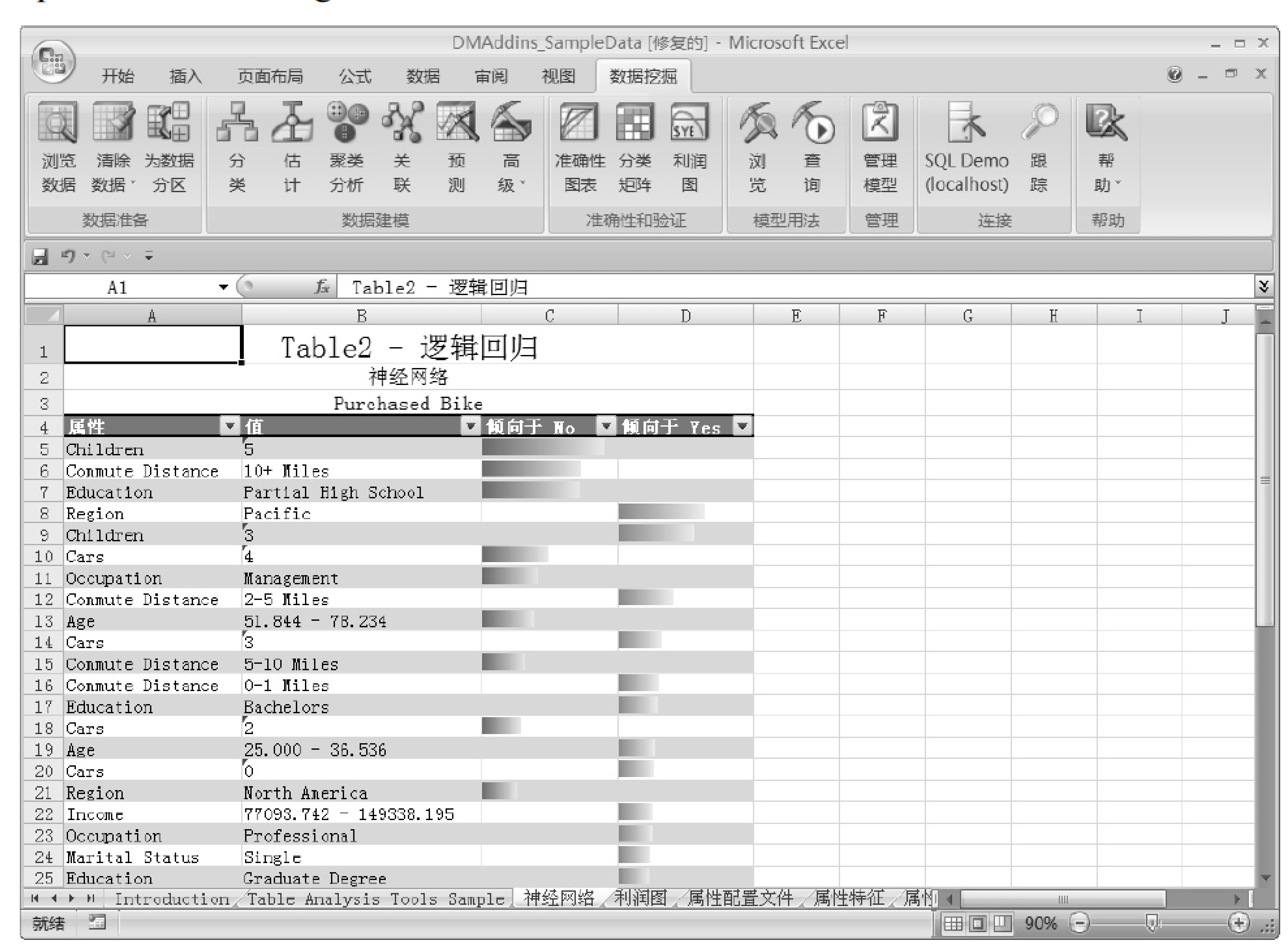


图 13-9 复制到 Excel

Step10：单击【数据挖掘】中的【准确性图表】按钮，弹出如图 13-10 所示的【准确性图表向导入门】窗口。

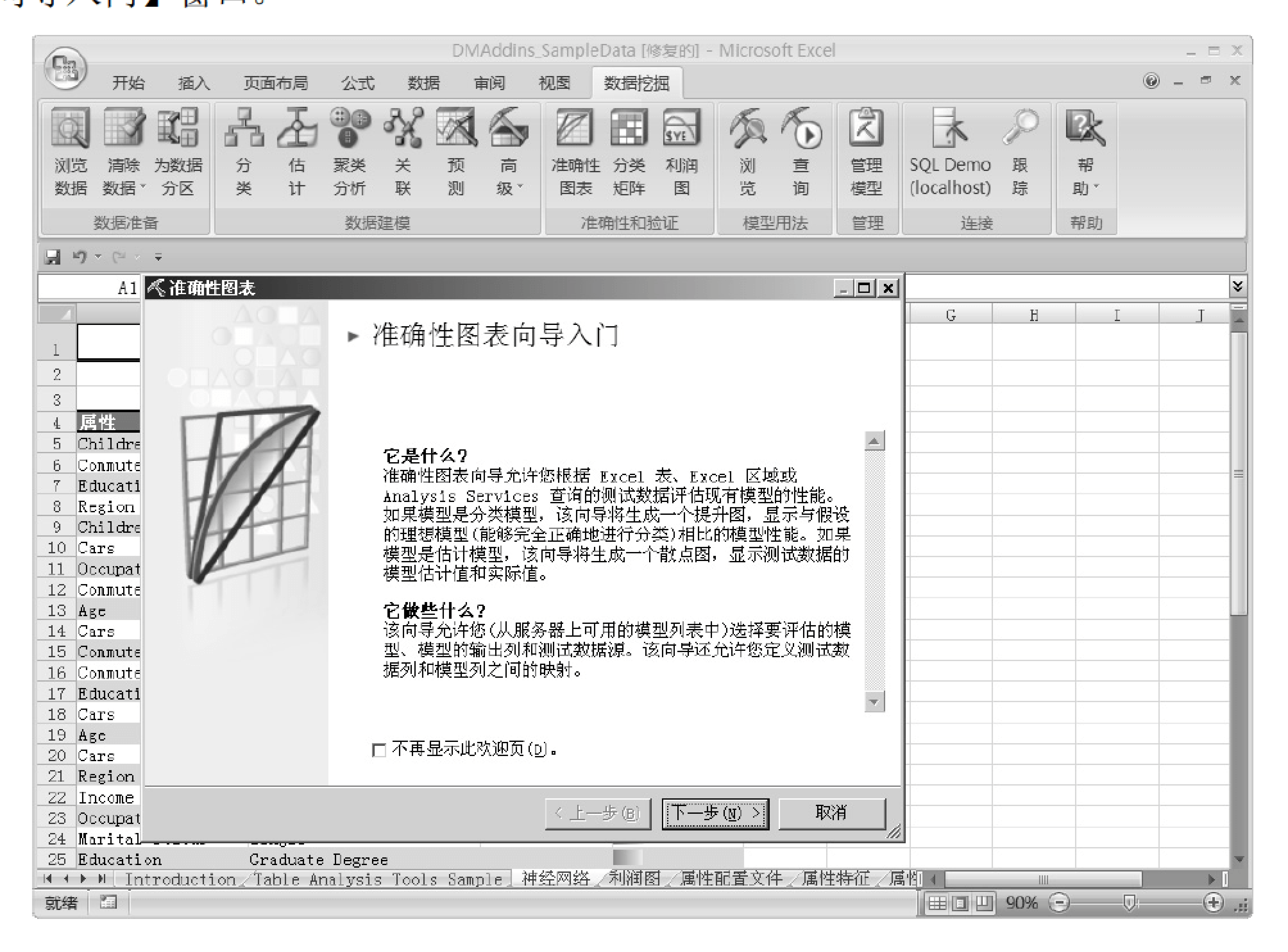


图 13-10 【准确性图表向导入门】窗口

Step11：在如图 13-11 所示的【选择模型】窗口中，单击【下一步】按钮。

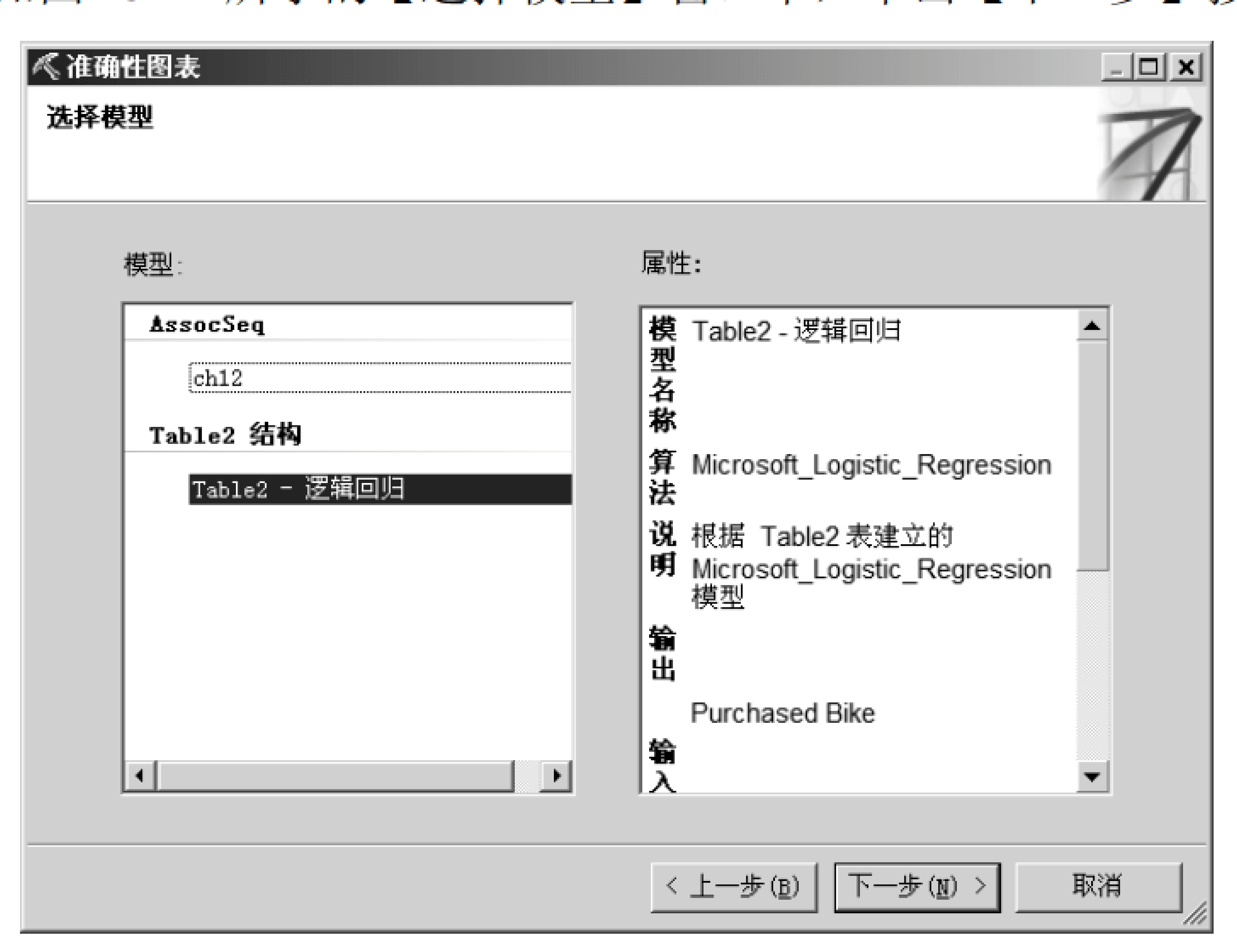


图 13-11 【选择模型】窗口

Step12：在【要预测的挖掘列】下拉列表框中选择 Purchased Bike，并单击【下一步】按钮，如图 13-12 所示。

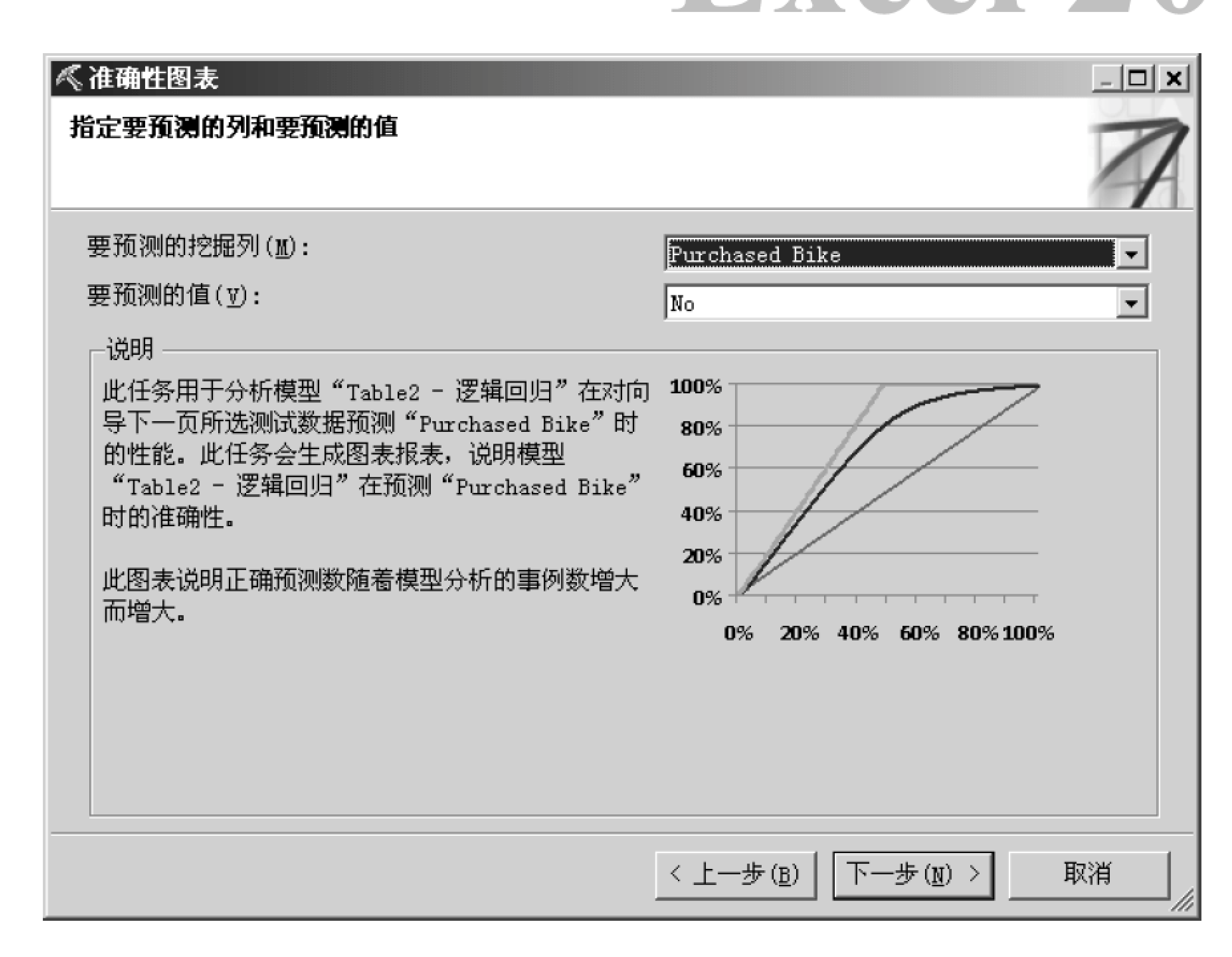


图 13-12　选择要预测的挖掘列

Step13：从【表】下拉列表框中选择数据表【'Table Analysis Tools Sample'!'Table2'】，并单击【下一步】按钮，如图 13-13 所示。



图 13-13　选择数据表

Step14：在如图 13-14 所示的【指定关系】窗口中，单击【完成】按钮。



图 13-14 【指定关系】窗口

Step15：得到模型“Table2-逻辑回归”的准确性图表如图 13-15 所示。

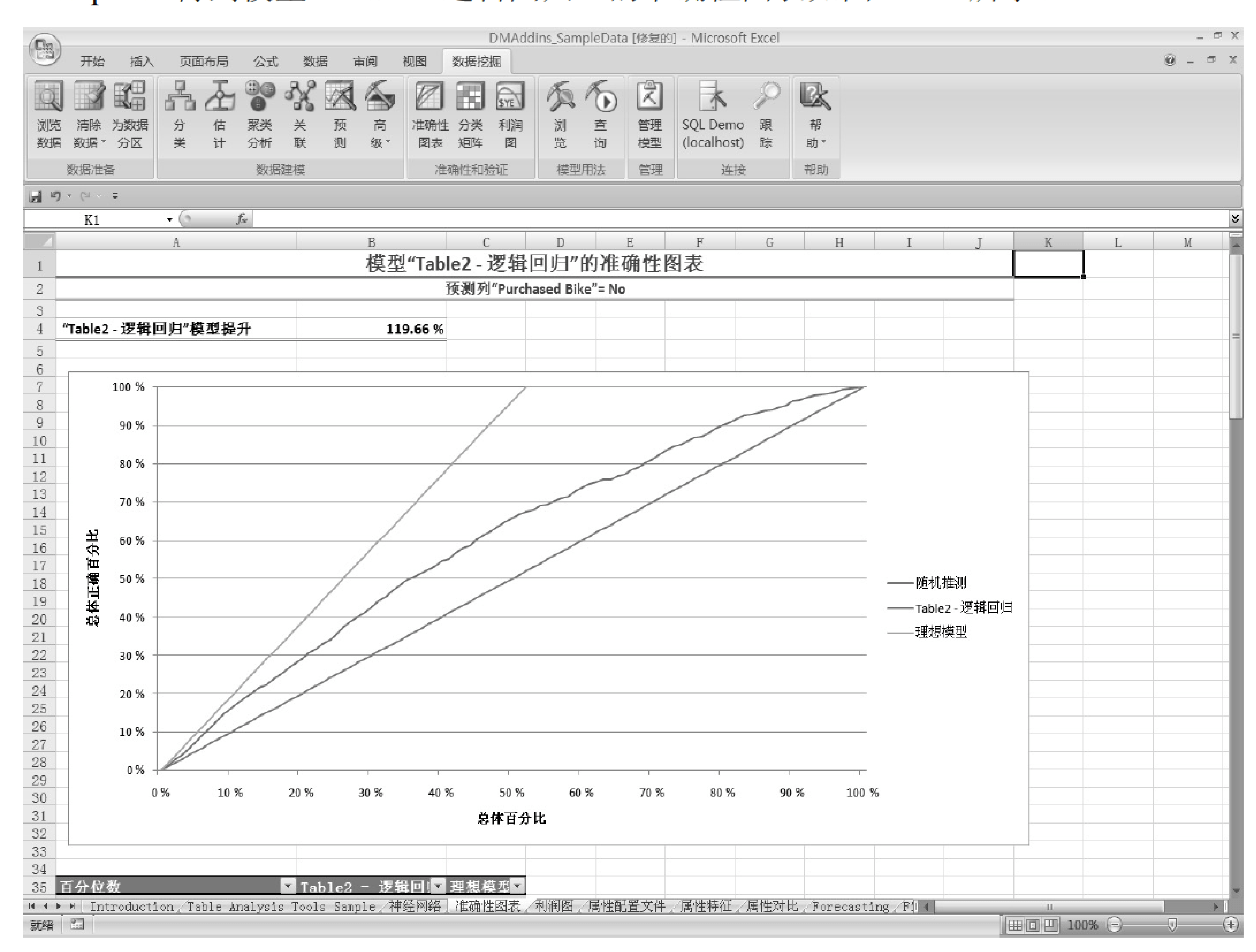


图 13-15 准确性图表

Step16：得到模型“Table2-逻辑回归”的准确性百分位数表如图 13-16 所示。

百分位数	Table2 - 逻辑回归	理想模型
0 %	0.00 %	0.00 %
1 %	1.35 %	1.93 %
2 %	2.89 %	3.85 %
3 %	4.05 %	5.78 %
4 %	5.78 %	7.71 %
5 %	7.51 %	9.63 %
6 %	9.44 %	11.56 %
7 %	11.18 %	13.49 %
8 %	12.91 %	15.41 %
9 %	14.84 %	17.34 %
10 %	16.18 %	19.27 %
11 %	17.53 %	21.19 %
12 %	19.08 %	23.12 %

图 13-16　准确性百分位数表

Step17：接着单击【数据挖掘】中的【分类矩阵】按钮，弹出如图 13-17 所示的【分类矩阵向导入门】窗口，单击【下一步】按钮。

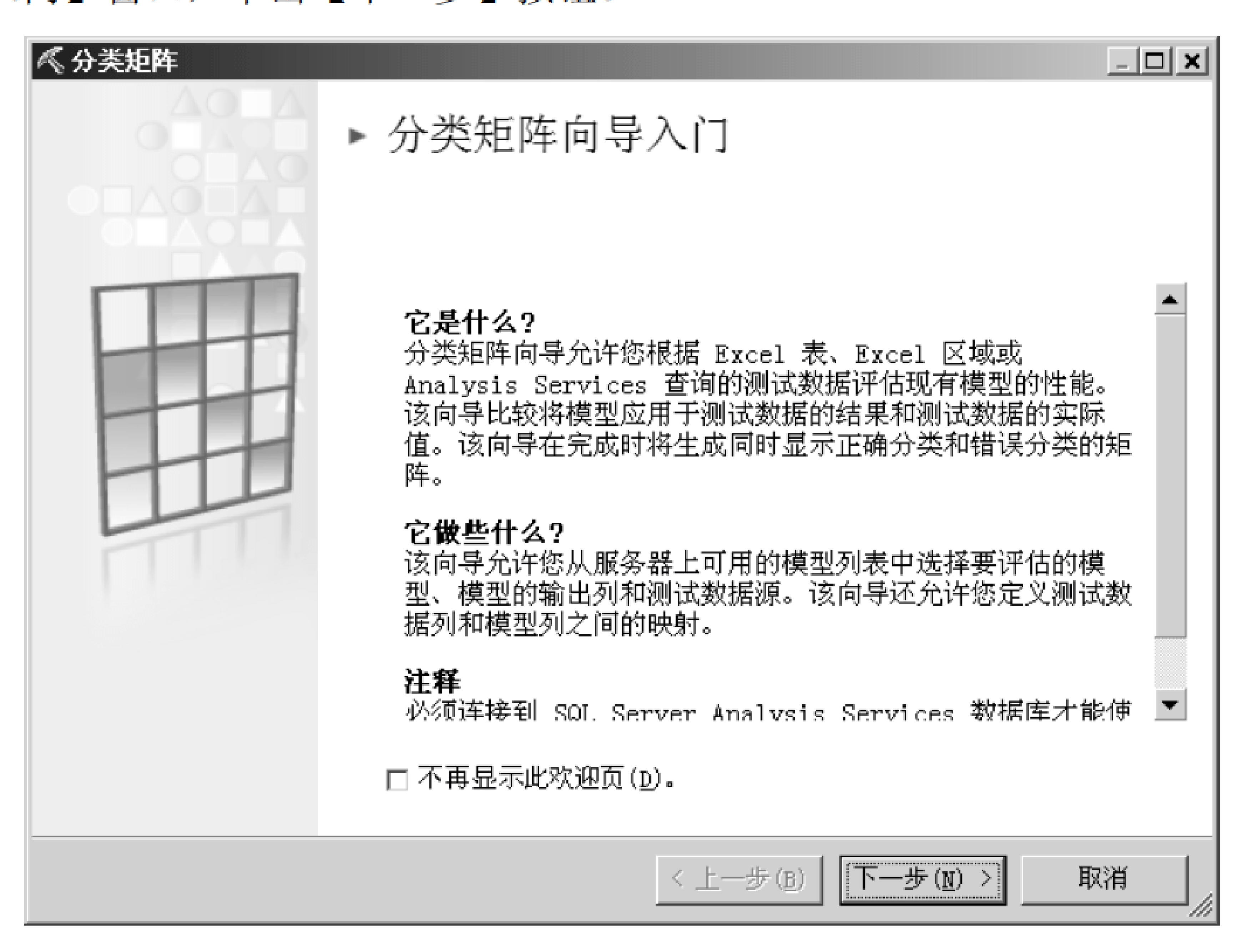


图 13-17　【分类矩阵向导入门】窗口

Step18：在如图 13-18 所示的【选择模型】窗口中，单击【下一步】按钮。



图 13-18 【选择模型】窗口

Step19：在【要预测的挖掘列】下拉列表框中选择 Purchased Bike，再单击【下一步】按钮，如图 13-19 所示。

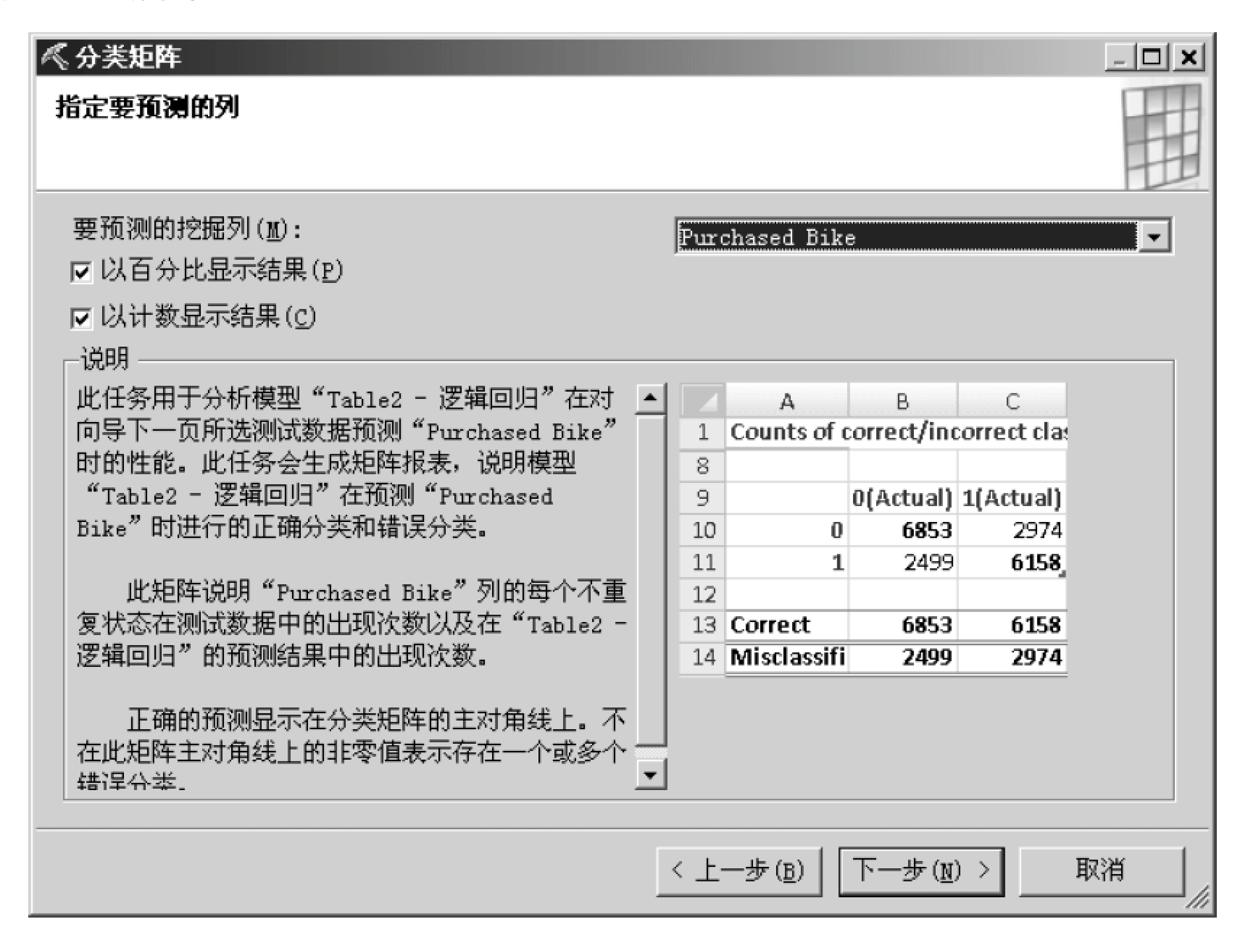


图 13-19 选择要预测的挖掘列

Step20：在【表】下拉列表中选择数据表【'Table Analysis Tools Sample'!'Table2'】，再单击【下一步】按钮，如图 13-20 所示。

Step21：在如图 13-21 所示的【指定关系】窗口中，单击【完成】按钮。

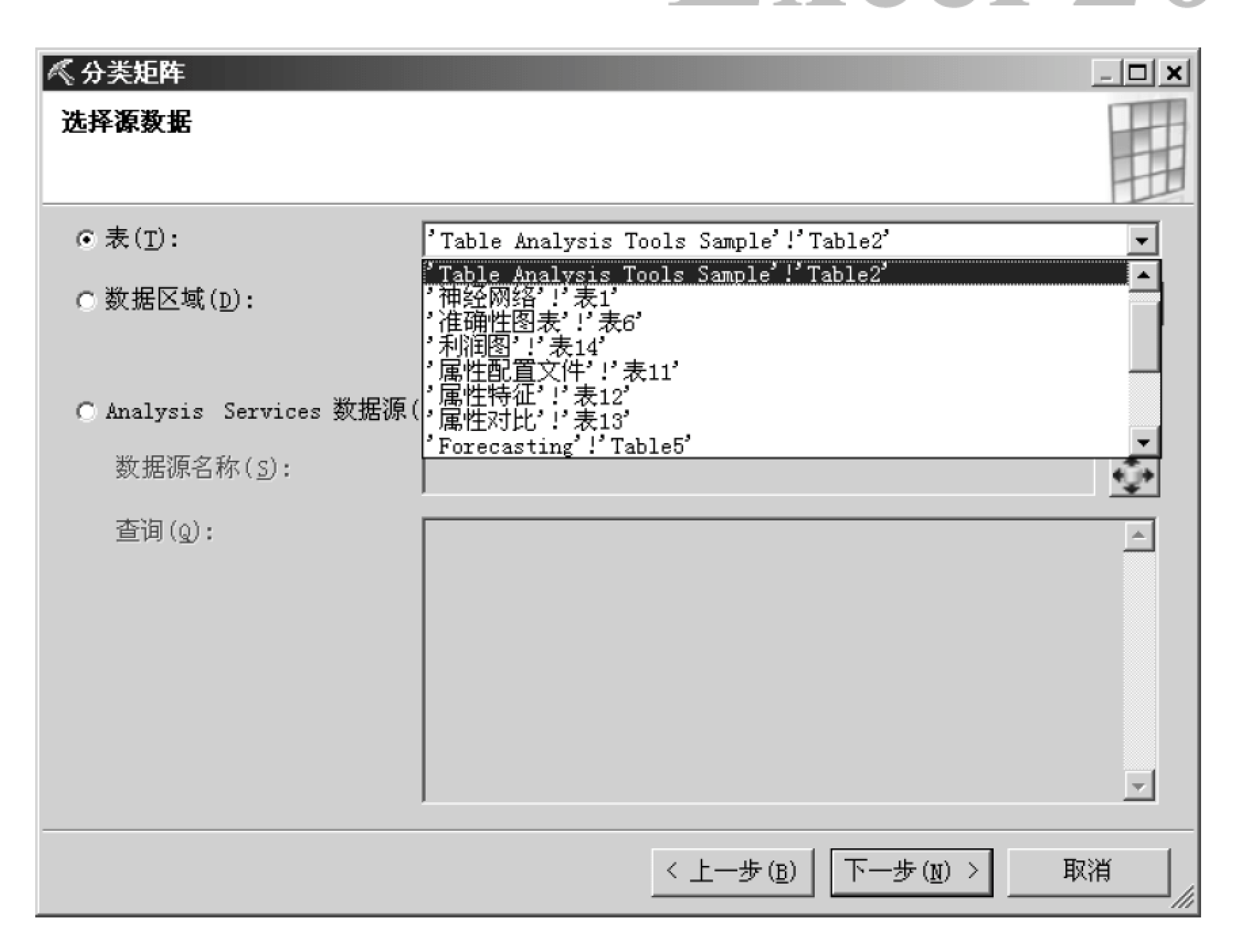


图 13-20 选择数据表



图 13-21 【指定关系】窗口

Step22：此时可得到分类矩阵，如图 13-22 所示。由表可知分类正确率达 66.10%、分类错误率为 33.90%。

Step23：单击【数据挖掘】中的【利润图】按钮，如图 13-23 所示。

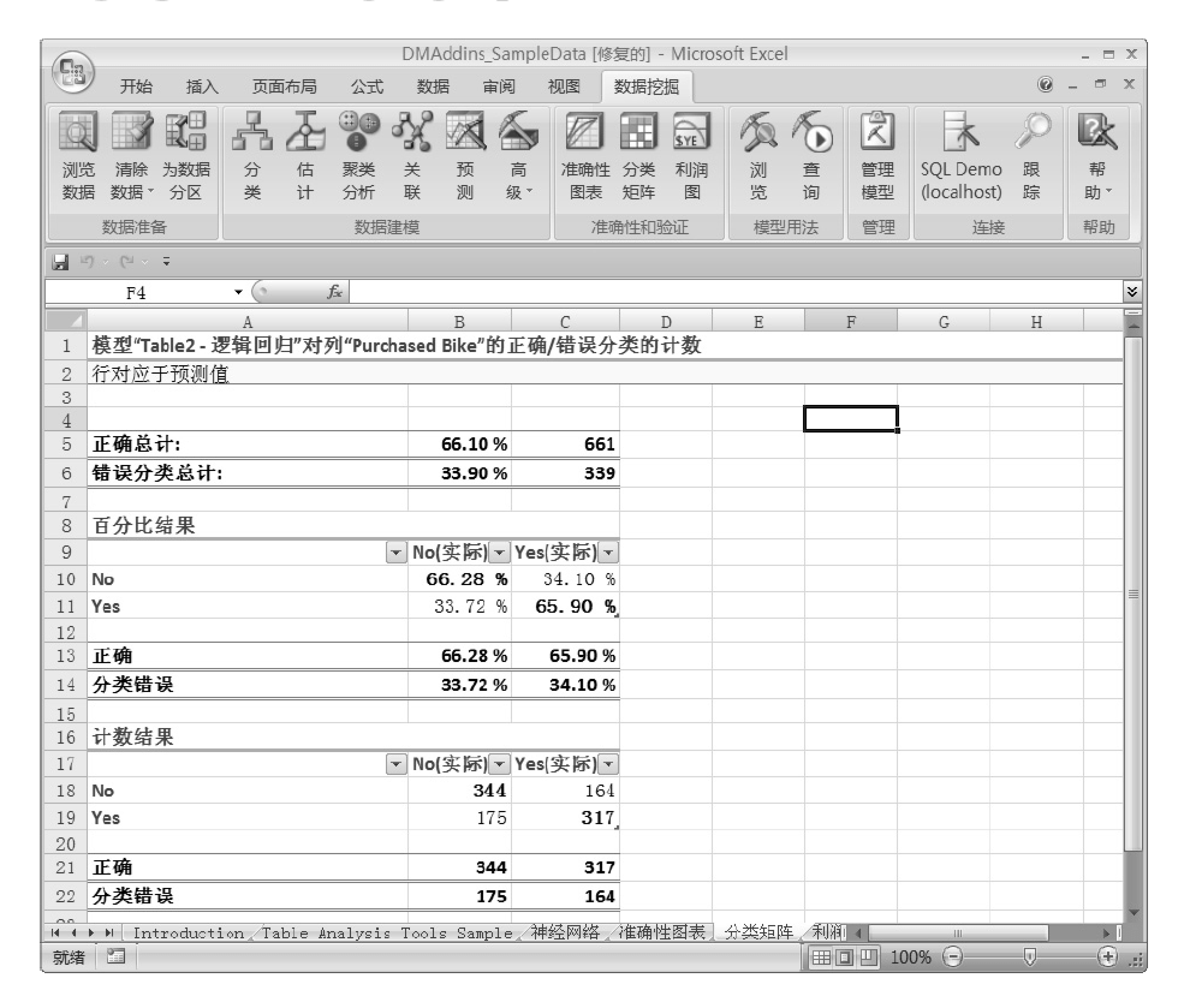


图 13-22　分类矩阵

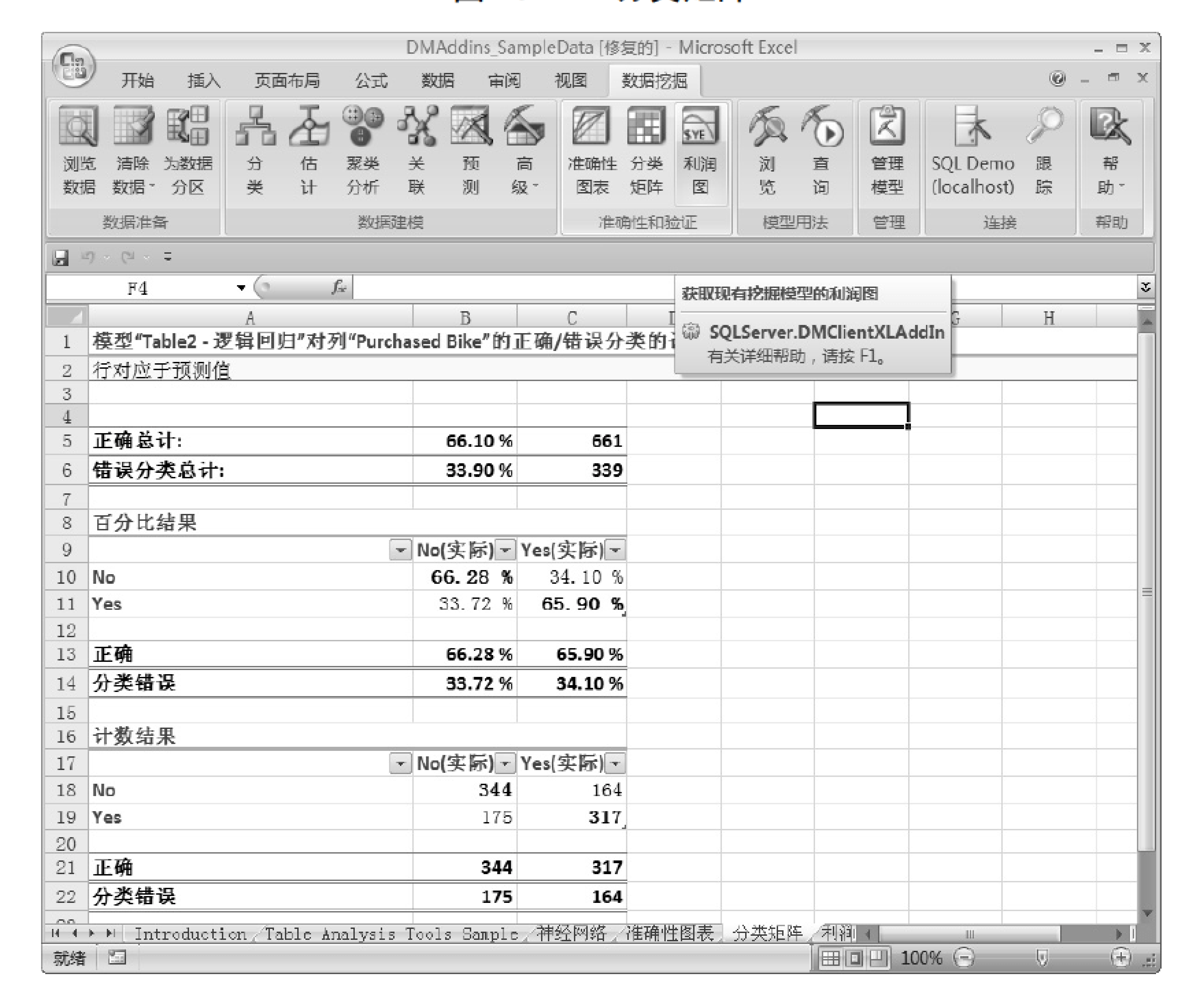


图 13-23　建立利润图



Step24：在如图 13-24 所示的【利润图向导入门】窗口中，单击【下一步】按钮。

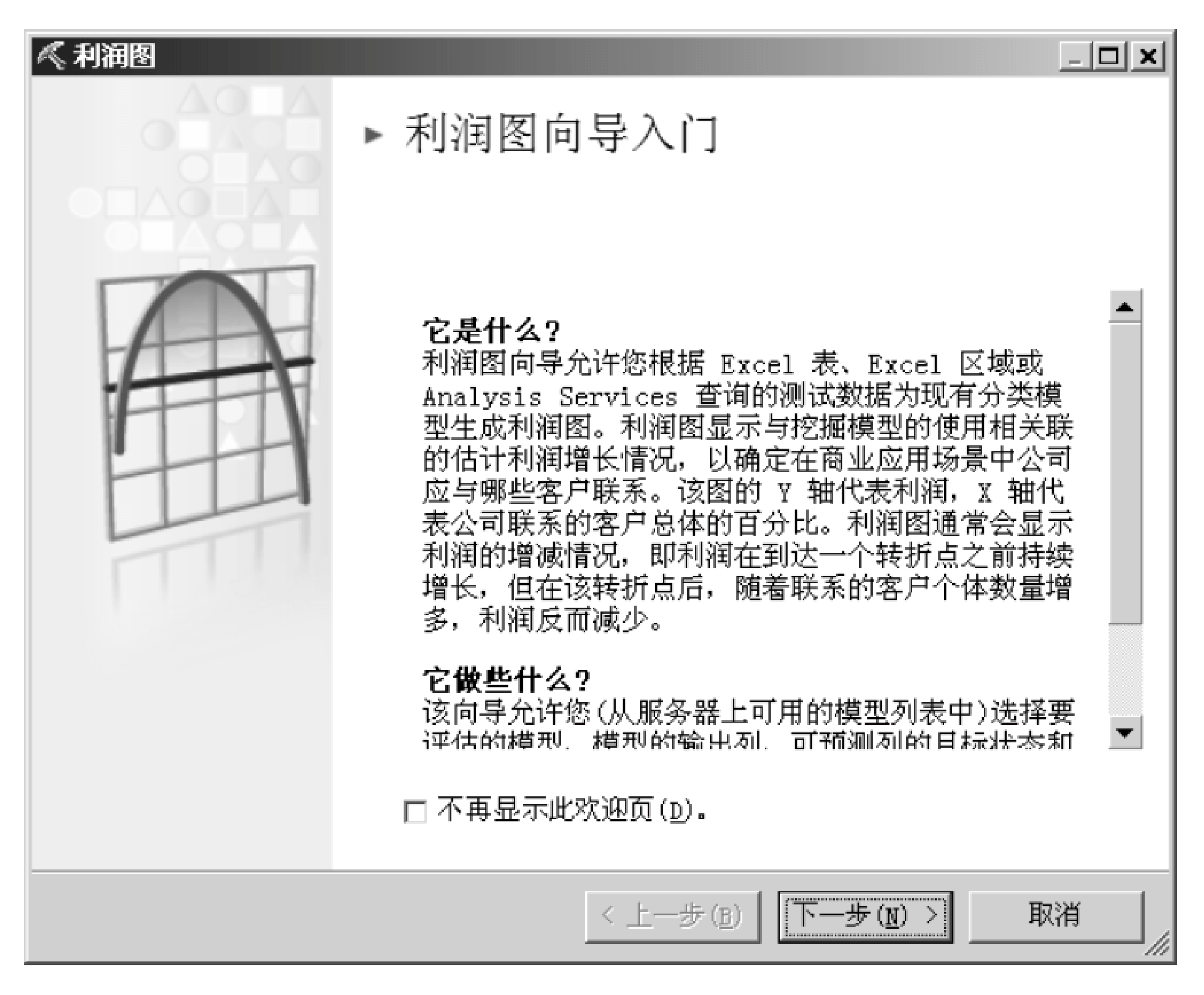


图 13-24 【利润图向导入门】窗口

Step25：在如图 13-25 所示的【选择模型】窗口中，单击【下一步】按钮。



图 13-25 【选择模型】窗口

Step26：在如图 13-26 所示的【指定利润图参数】窗口中，设定要预测的挖掘列、要预测的值、目标总体、固定成本、单项成本、单项收入，单击【下一步】按钮。

Step27：在【表】下拉列表框中选择数据表【'Table Analysis Tools Sample'!'Table2'】，

如图 13-27 所示，单击【下一步】按钮。

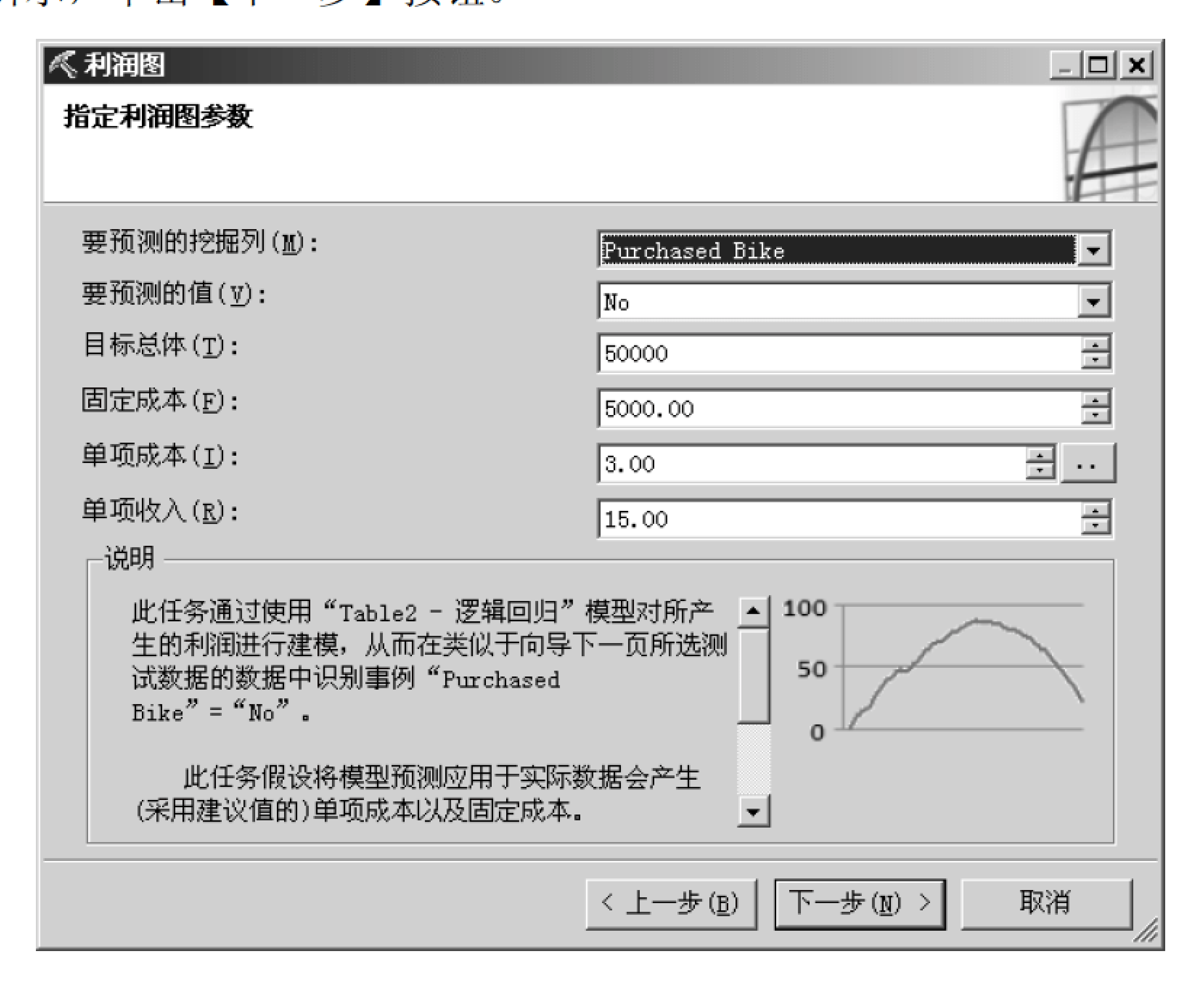


图 13-26　【指定利润图参数】窗口



图 13-27　选择数据表

Step28：在如图 13-28 所示的【指定关系】窗口中，单击【完成】按钮。

Step29：得到利润图和利润百分位数表，如图 13-29、图 13-30 所示。



图 13-28 【指定关系】窗口

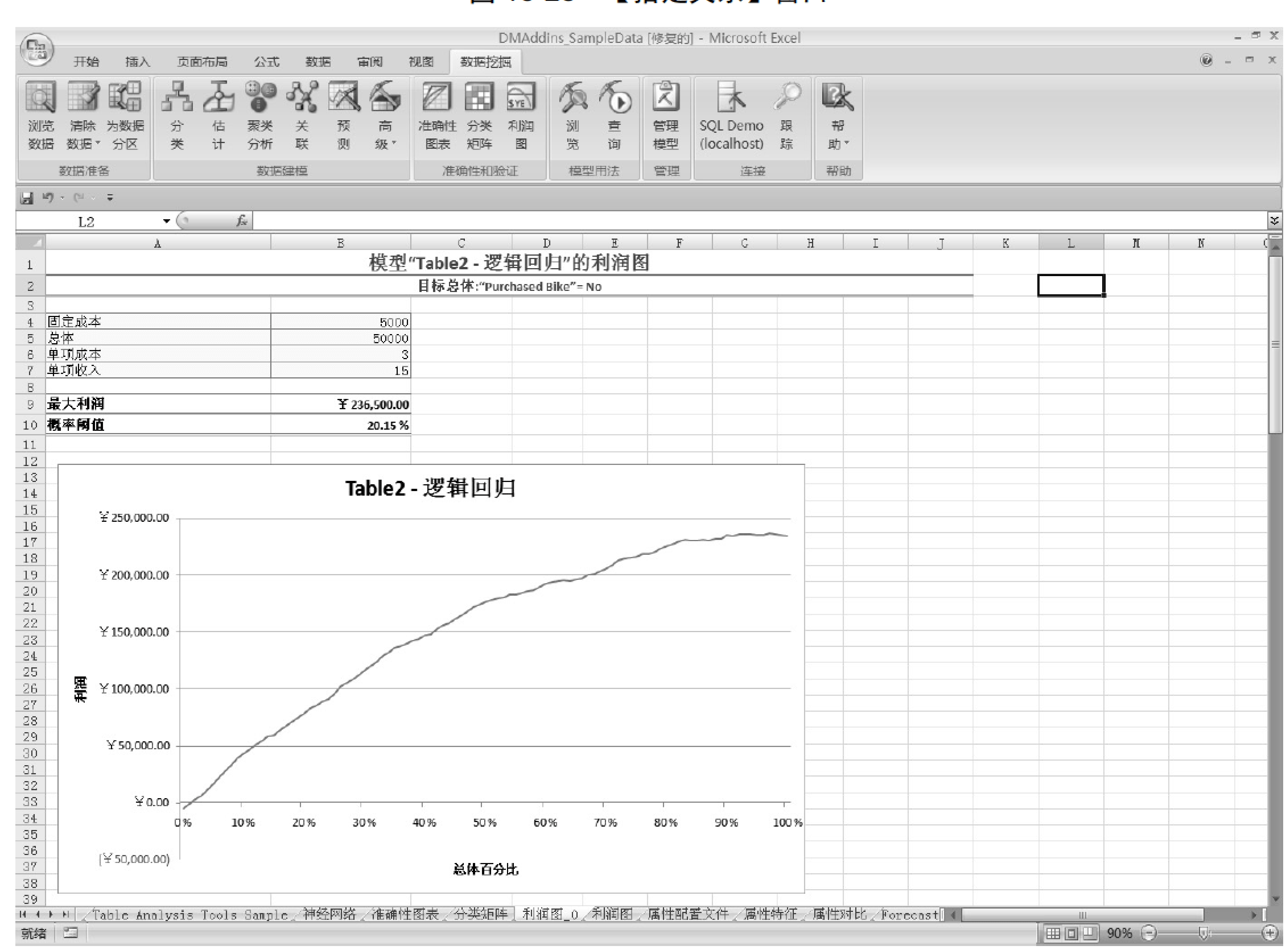
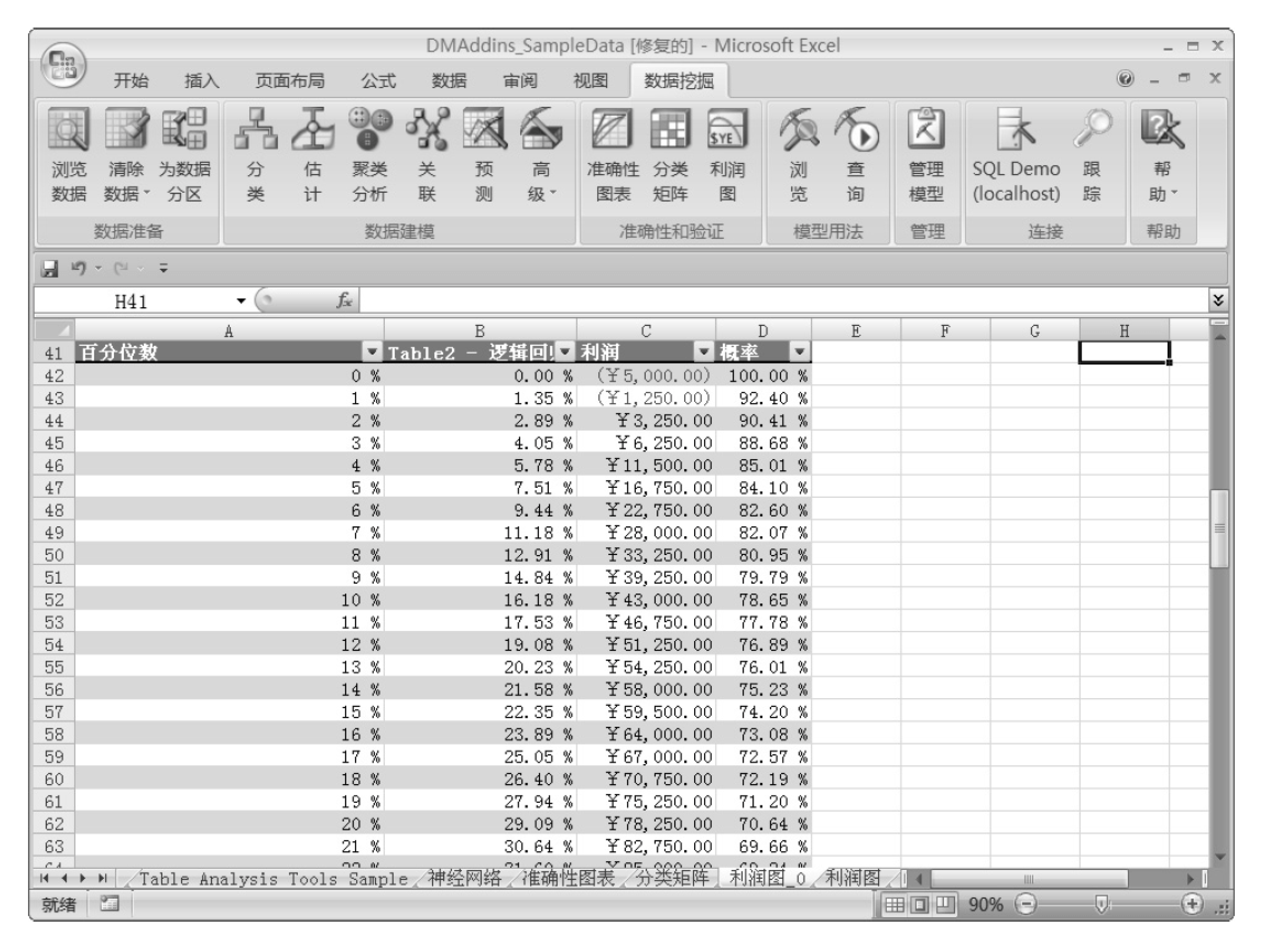


图 13-29 利润图



百分位数	Table2 - 逻辑回!	利润	概率
0 %	0.00 %	(¥5,000.00)	100.00 %
1 %	1.35 %	(¥1,250.00)	92.40 %
2 %	2.89 %	¥3,250.00	90.41 %
3 %	4.05 %	¥6,250.00	88.68 %
4 %	5.78 %	¥11,500.00	85.01 %
5 %	7.51 %	¥16,750.00	84.10 %
6 %	9.44 %	¥22,750.00	82.60 %
7 %	11.18 %	¥28,000.00	82.07 %
8 %	12.91 %	¥33,250.00	80.95 %
9 %	14.84 %	¥39,250.00	79.79 %
10 %	16.18 %	¥43,000.00	78.65 %
11 %	17.53 %	¥46,750.00	77.78 %
12 %	19.08 %	¥51,250.00	76.89 %
13 %	20.23 %	¥54,250.00	76.01 %
14 %	21.58 %	¥58,000.00	75.23 %
15 %	22.35 %	¥59,500.00	74.20 %
16 %	23.89 %	¥64,000.00	73.08 %
17 %	25.05 %	¥67,000.00	72.57 %
18 %	26.40 %	¥70,750.00	72.19 %
19 %	27.94 %	¥75,250.00	71.20 %
20 %	29.09 %	¥78,250.00	70.64 %
21 %	30.64 %	¥82,750.00	69.66 %

图 13-30 利润百分位数表

第 14 章　类神经网络

14.1　基 本 概 念

为了在语音及影像识别领域实现与人脑相似的功能，自 1940 年起，科学家开始了类神经网络方面的研究，模仿最简单的神经元模型，构建最原始的类神经网络（artificial neural network，ANN）。历经 40 年的发展，类神经网络的研究工作与生理学、心理学、计算机科学等学科交叉渗透，成为新的研究领域。

一部机器的运作或一个事件的发生常常有相对应的因果关系（例如：打开电器的开关，电器开始工作；脚踩油门，汽车加速），将打开开关与脚踩油门的动作称为系统的输入，电器与汽车称为系统，而电器的运作与汽车的速度称为系统的输出，整个输入与输出的关系可以用一个方块图来表示，如图 14-1 所示。



图 14-1　系统的输入与输出关系

类神经网络的一个优点在于无须了解系统的数学模型的具体形式，而直接用神经网络取代系统的模型，一样可以得到输入与输出之间的对应关系。其方块图如图 14-2 所示。

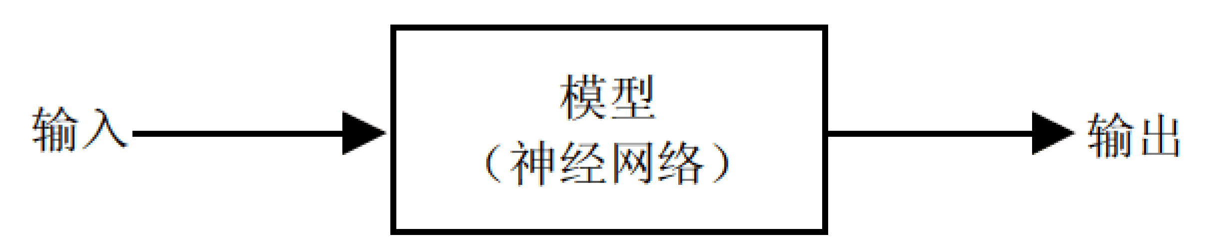


图 14-2　神经网络的输入与输出关系

人类的大脑大约由 10^{11} 个神经细胞（nerve cells）组成，而每个神经细胞又有 10^{4} 个突触（synapses）与其他细胞互相连结成一个非常复杂的神经网络。一个神经单元由一个细胞主体（cell body）构成，而细胞主体则有一些分支凸起的树状突起（dendrite）和一个单一分支的轴突（axom）。树状突起接收其他的神经单元的信号，而当其所接受的脉动（impulse）信号超过某一特定的阈值（threshold），这个神经单元就会被点燃（fire），并产生一个脉动传递到轴突。

在轴突末端的分支称为胞突缠络（synapse），它是神经与神经的连络点，它可以是抑制的或者是刺激的。抑制的胞突缠络会降低所传送的脉冲；刺激的胞突缠络则会加强所传送的脉冲。当外界刺激由神经细胞传递到大脑，大脑便会将相应指令传递至相关的受动器（effectors）做出反应（例如：手的皮肤接触到烫的物体手会立即放开），适当的反应往往需要经过反复的训练和记忆才能实现。如果大脑受到损害（例如中风患者），便需要借助康

复的方式，重新学习。

图 14-3 描述了一个类神经元模型。

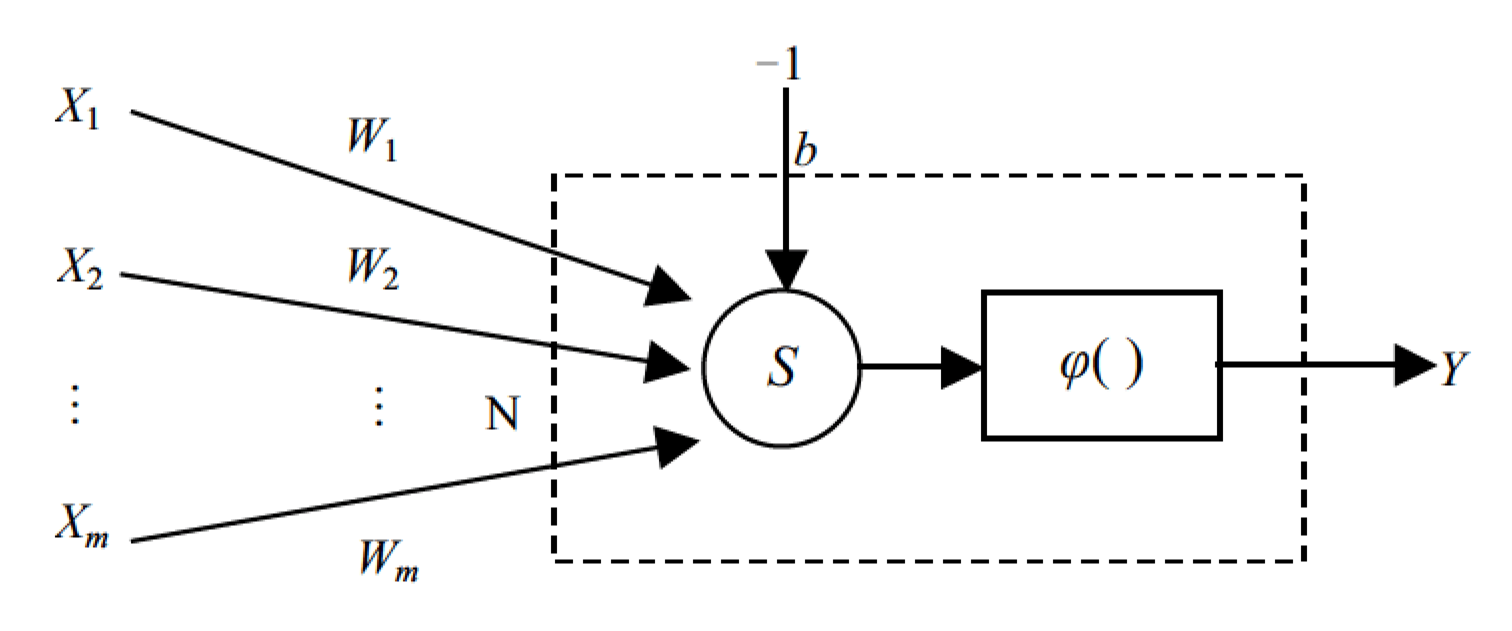


图 14-3　类神经元模型

其中：

X 称为神经元的输入（input）。

W 称为连接权数（weights）。

b 称为阈值（bias），有偏移的效果。

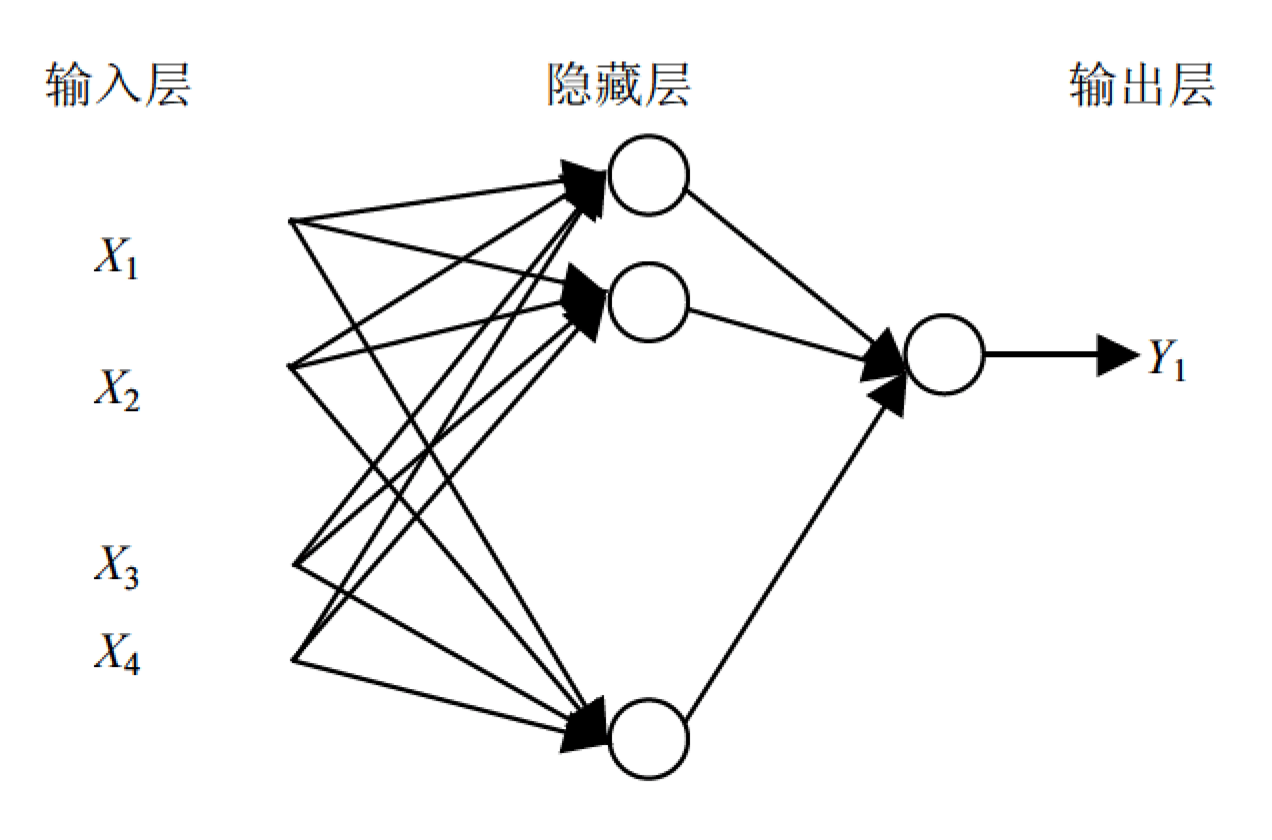
S 称为集成单元（summation），此部分是将每一个输入与连接权数相乘后做一加总的动作。

$\varphi(\ )$ 称为活化函数（activation function），通常是非线性函数，有数种不同的形式，其目的是将 S 的值映射到所需的输出；

Y 称为输出（output），即所需的结果。

虚线的部分是类神经元，类神经网络的训练就是调整连接权数，使其变得更大或是更小，通常由随机的方式产生介于+1～−1 之间的初始值。连接权数可视为一种加权效果，其值越大，代表连结的神经元越容易被激发，对类神经网络的影响也更大；反之，代表该输入对类神经网络并无太大的影响，而太小的连接权数通常可以移除以节省计算时间与储存空间。

图 14-4 显示的是四个输入与一个输出的反向传递网络模型。



*圆圈代表神经元。

图 14-4　反向传递网络模型

这个网络由三层类神经单元所组成。第一层是由输入单元所组成的输入层，这些输入单元通过固定强度的链接连接到特征检测单元后，再通过可调整强度的链接连接到输出层中的输出单元，最后，每个输出单元对应到某一种特定的分类。调整链接强度的过程就是机器学习的过程。

14.2 类神经网络的架构与训练算法

类神经网络的架构与训练算法如表 14-1 所示。

表 14-1 类神经网络架构与训练算法

架　构	训 练 算 法
单层网络	可形成两个决定区域（decision region），而此二区域由一超平面（hyperplane）加以分隔开来。有一特殊情形就是，若网络只涉及两个输入，则超平面便退化成一条直线
多层网络	在输入层节点与输出层节点间多了一层或多层的隐藏层（hidden layer），即输入节点没有直接接往输出节点

14.3 类神经网络的特性

类神经网络具有以下几种特性，如表 14-2 所示。

表 14-2 类神经网络的特性

特　性	说　明
平行处理	随着超大型平行处理的发展，成为人工智能中最活跃的研究领域
容错性（fault tolerance）	在操作上具有很高的容错度，整个神经网络都会参与解决问题的运作。即使 10%的神经网络失效，仍能照常运作
结合记忆特性（associative memory）	又称内容寻址记忆（content addressable memory），可以记忆曾经训练过的输入样式以及对应的理想输出值
优化（optimization）	处理非算法表示的问题，算法密集型的问题
超大规模集成电路实现（VLSI implementation）	神经网络的结构具有高度的互相连接（interconnection），而且简单、有规则性（regularity），容易用超大规模集成电路（VLSI）来完成

14.4 类神经网络应用

由于类神经网络对于输入映射到输出有记忆与学习的功能，并且对缺失的输入有推断的功能，因此类神经网络可运用于各种领域中，举例如表 14-3 所示。

表 14-3　类神经网络应用

领　　域	具 体 应 用
工业应用	❑ 控制器设计与系统鉴别 ❑ 产品质量分析（例：汽水瓶装盖与填充监测、珍珠分级） ❑ 机电设备诊断（例：数值电路诊断、模拟 IC 诊断、汽车引擎诊断） ❑ 化工程序诊断（例：化工厂制作程序故障诊断） ❑ 实验数据模型建立（例：复合材料行为模型建立） ❑ 工程分析与设计（例：钢梁结构、道路铺面状况评级）
商业应用	❑ 股票投资（例：大盘基本分析、大盘技术分析、个股技术分析） ❑ 债券投资（例：债券分级、美国国库券利率预测） ❑ 期货、期权、外汇投资（例：期货投资、期权投资、外汇投资） ❑ 商业信用评估（例：贷款信用审核、信用卡信用审核） ❑ 其他商业应用（例：直销顾客筛选、不动产鉴价）
管理应用	❑ 策略管理（例：市场需求预测方法的选择、雇工人数规划） ❑ 时程管理（例：排程策略选择、工作排程） ❑ 质量管理（例：管制图判读、半导体制造过程所需蚀刻时间估计）
信息应用	❑ 影像辨识系统（例：指纹识别、卫星遥测影像分析、医学影像识别） ❑ 信号分类 ❑ 其他信息应用（例：雷达信号分类、声纳信号分类）
科学应用	❑ 医学（例：皮肤病诊断、头痛疾病诊断、心脏病诊断、基因分类） ❑ 化学（例：化合物化学结构识别、蛋白质结构分析） ❑ 其他科学应用（例：体操选手运动伤害分析、时间序列分析方法选择）
其他领域的应用	❑ 函数模型构建（例：自来水厂水质处理操作） ❑ 预测模型构建（例：电力负载预测、太阳黑子活动预测） ❑ 决策模型构建（例：排程策略选择、建筑结构材料选择）

14.5　类神经网络优缺点

类神经网络的优点有：

① 类神经网络可以构建非线性的模型，模型的准确度高。

② 类神经网络有良好的推广性，对于缺失的输入也可推断得到正确的输出。

③ 类神经网络可以接受离散或连续变量作为输入，适应性强。

④ 类神经网络可应用的领域广泛，建模能力强。

⑤ 类神经网络具有模糊推论能力，允许输入变量具有模糊性，归纳学习比较难具备这一能力。

但其缺点也是明显的：

① 类神经网络因为其中间变量（即隐藏层）可以是一层或二层，数目也可设为任意数

目，而且有学习速率等参数需要设定，工作相当费时。

② 类神经网络用迭代方式更新键结值与阈值，计算量大，相当耗费计算机资源。

③ 类神经网络的解有无限多组，无法得知哪一组的解为最佳解。

④ 类神经网络训练的过程中无法得知需要多少神经元个数，太多或太少的神经元均会影响系统的准确性，因此往往需以试误的方式得到适当的神经元个数。

⑤ 类神经网络因为是以建立数值结构（含加权值的网络）来学习，其知识结构是隐性的，缺乏解释能力；而归纳学习以建立符号结构（如决策树）来学习，其知识结构是显性的，具有解释能力。

⑥ 类神经网络并非人脑。人脑有更复杂的结构，不仅能调整连结强度的大小，还可以建立新的连结。

⑦ 类神经网络目前仍不能仿真高度抽象的表示方式，例如符号。因此可能具有很差的抽象程度，它本身可能无法来描述高层次的程序。

⑧ 人类的某些智慧行为并不是平行的。许多高层次的推理行为在本质上似乎是有顺序的。

⑨ 人脑是一个相当大的组织，它具有上亿个神经。虽然在较小的系统中已确定可以达成一些有用的行为，但是具有更加智能的程序所需的神经个数，可能远超过实际能制作在计算机上的数目。

虽然有这些困难，但目前计算机的速度越来越快，类神经网络的训练时间可以更为缩短，相信在未来类神经网络的应用领域将会更为广泛，类神经网络具有相当的发展潜力，而且将成为研究的一个重要焦点。

14.6 Excel 2007 类神经网络

Microsoft 类神经网络算法使用迭代方法，将多层网络的参数优化，来预测多个属性。它可用于类别属性的分类以及连续属性的回归。

Step1：数据来源为 Microsoft 示例数据集，为 2002—2007 年自行车购买的数据集，建立类神经网络模型。单击【数据挖掘】下的【高级】按钮，弹出如图 14-5 所示的【创建模型向导入门】窗口，开始建立数据挖掘模型，单击【下一步】按钮。

Step2：选择挖掘算法的步骤，在【算法】下拉列表框中选择 Microsoft 神经网络，单击【下一步】按钮，如图 14-6 所示。

Step3：在选择数据行的步骤时，在各个变量后方有一栏是使用方式选择，用户可以选择一个变量的使用方式，包括输入、仅预测、输入和预测、键以及不使用等。本次使用是否购买自行车（Purchased Bike）作为预测变量 *Y*，其余变量作为解释变量建立模型，接着单击【下一步】按钮，如图 14-7 所示。

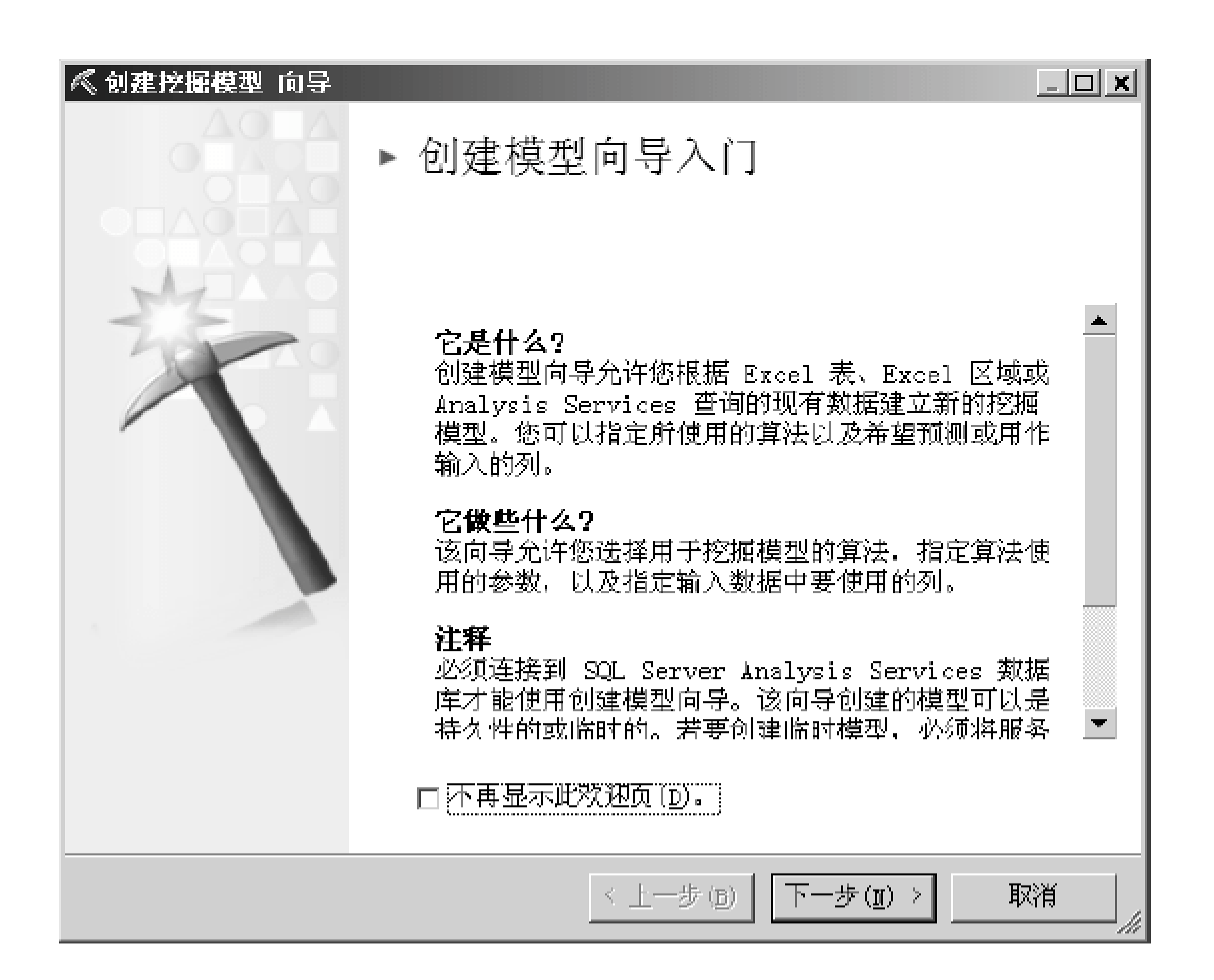


图 14-5 【创建模型向导入门】窗口

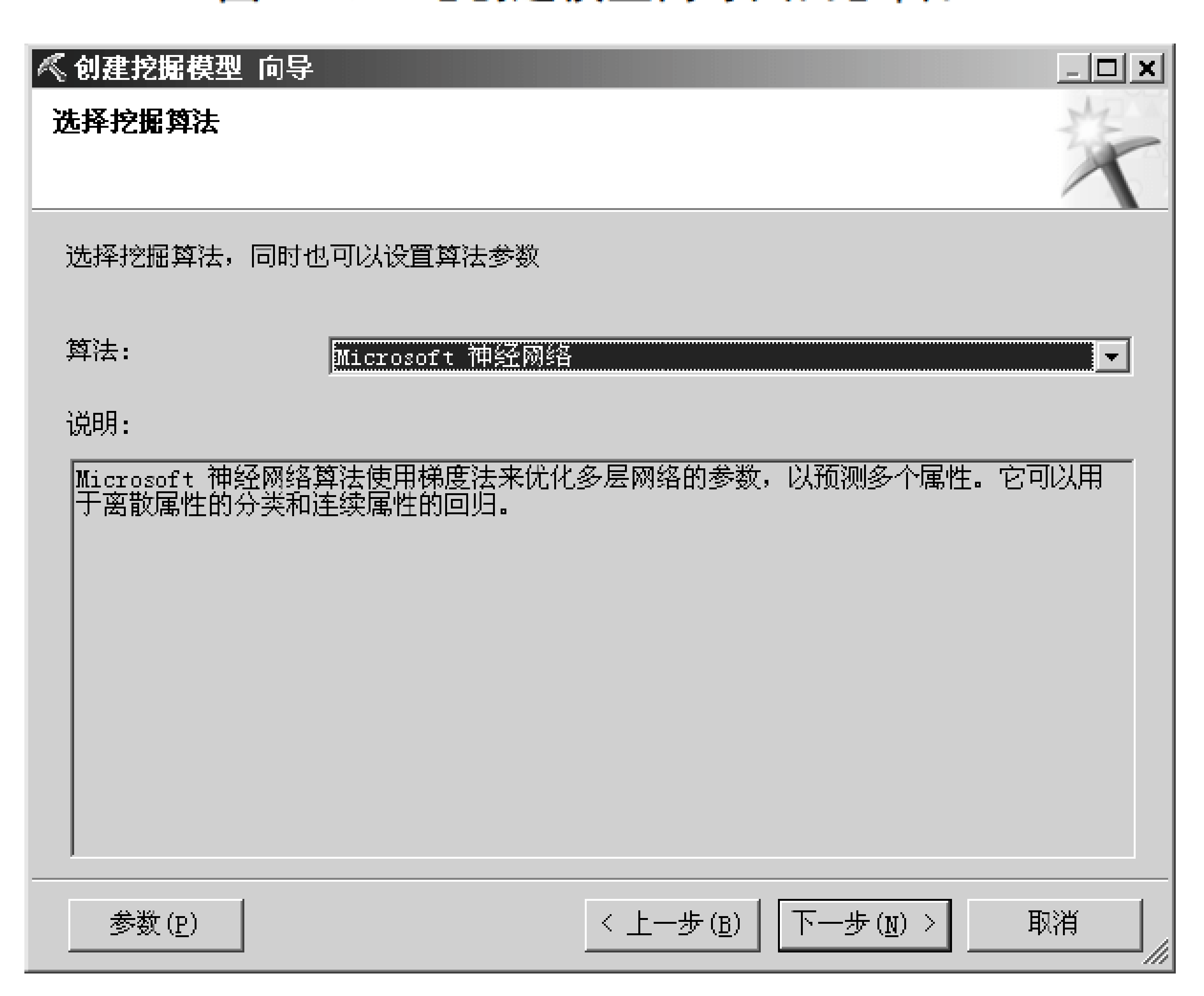


图 14-6 选择挖掘算法



图 14-7 选择列

Step4：在如图 14-8 所示的【完成】窗口中，单击【完成】按钮，开始构建数据挖掘模型。



图 14-8 【完成】窗口

Step5：弹出如图 14-9 所示的【浏览】窗口。可以利用变量属性来了解购买与不购买在变量属性上的差异。

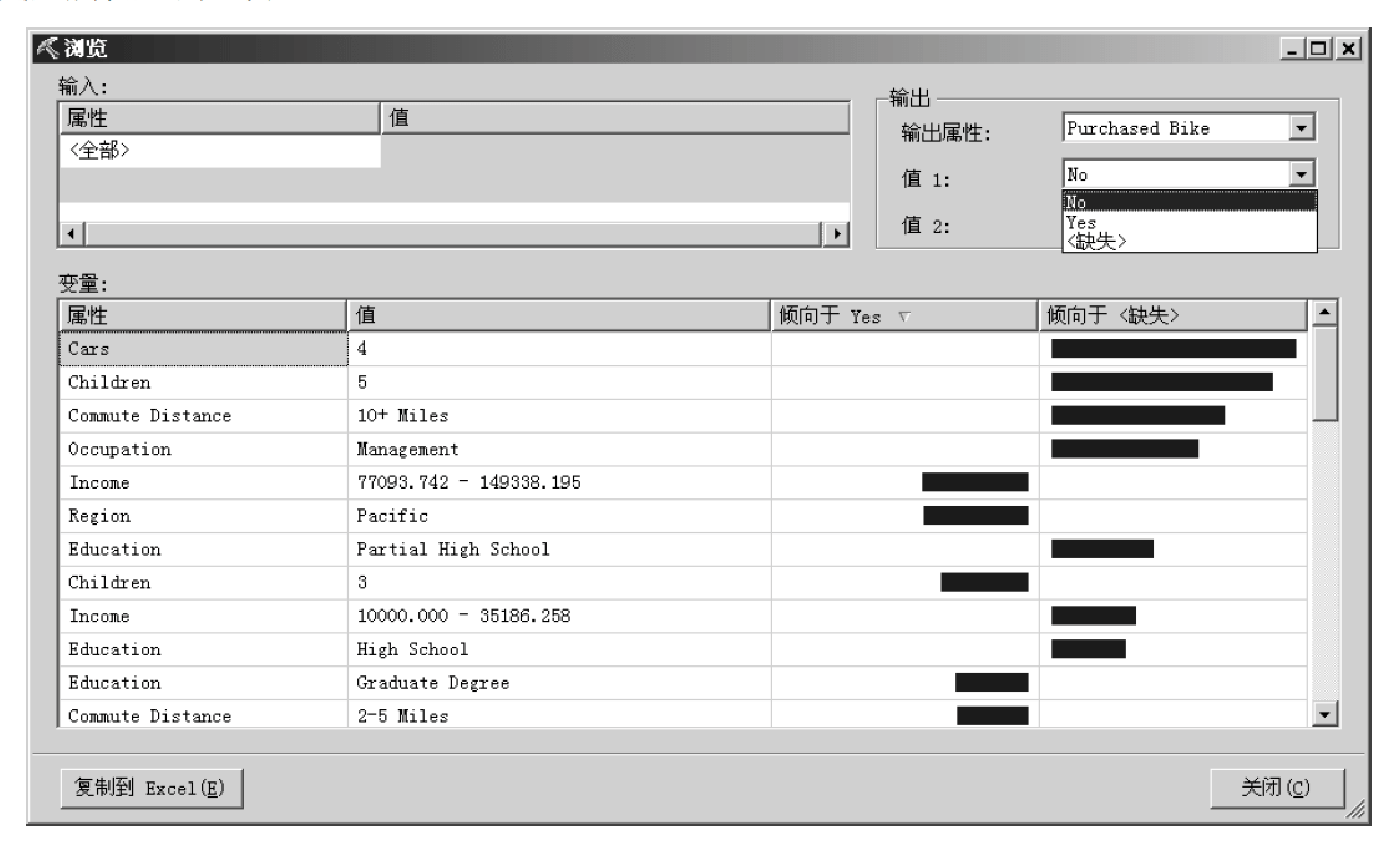


图 14-9 【浏览】窗口

Step6：利用图 14-10 可以比较购买或不购买的两个群体之间变量属性的差异，如果想将窗口复制到 Excel 窗口下操作，可以单击【复制到 Excel】按钮。

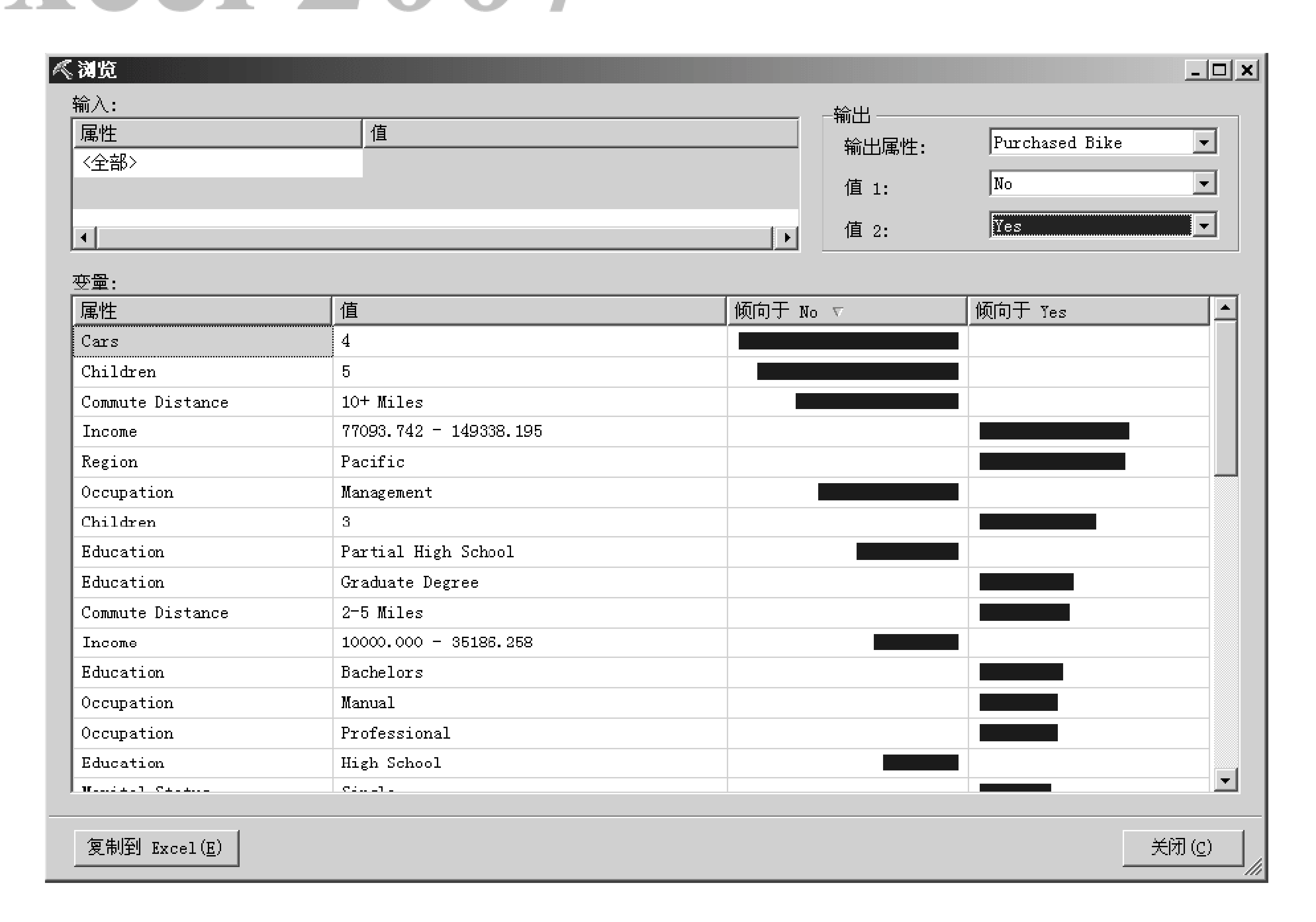


图 14-10　比较变量属性的差异

Step7：将图表复制到 Excel 中，如图 14-11 所示。

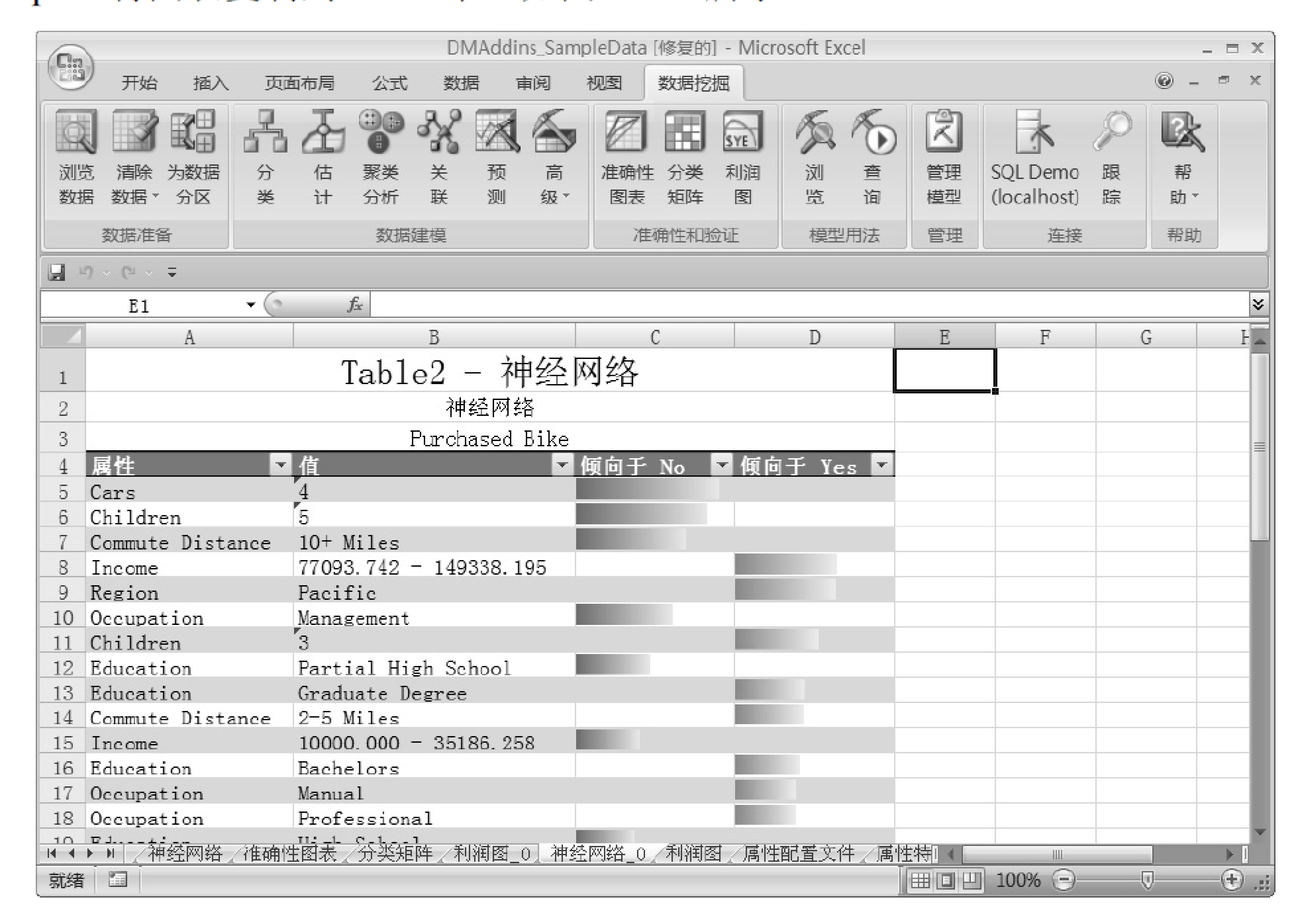


图 14-11　复制到 Excel

Step8：单击【数据挖掘】中的【准确性图表】按钮，弹出如图 14-12 所示的【准确性图表向导入门】窗口，单击【下一步】按钮。

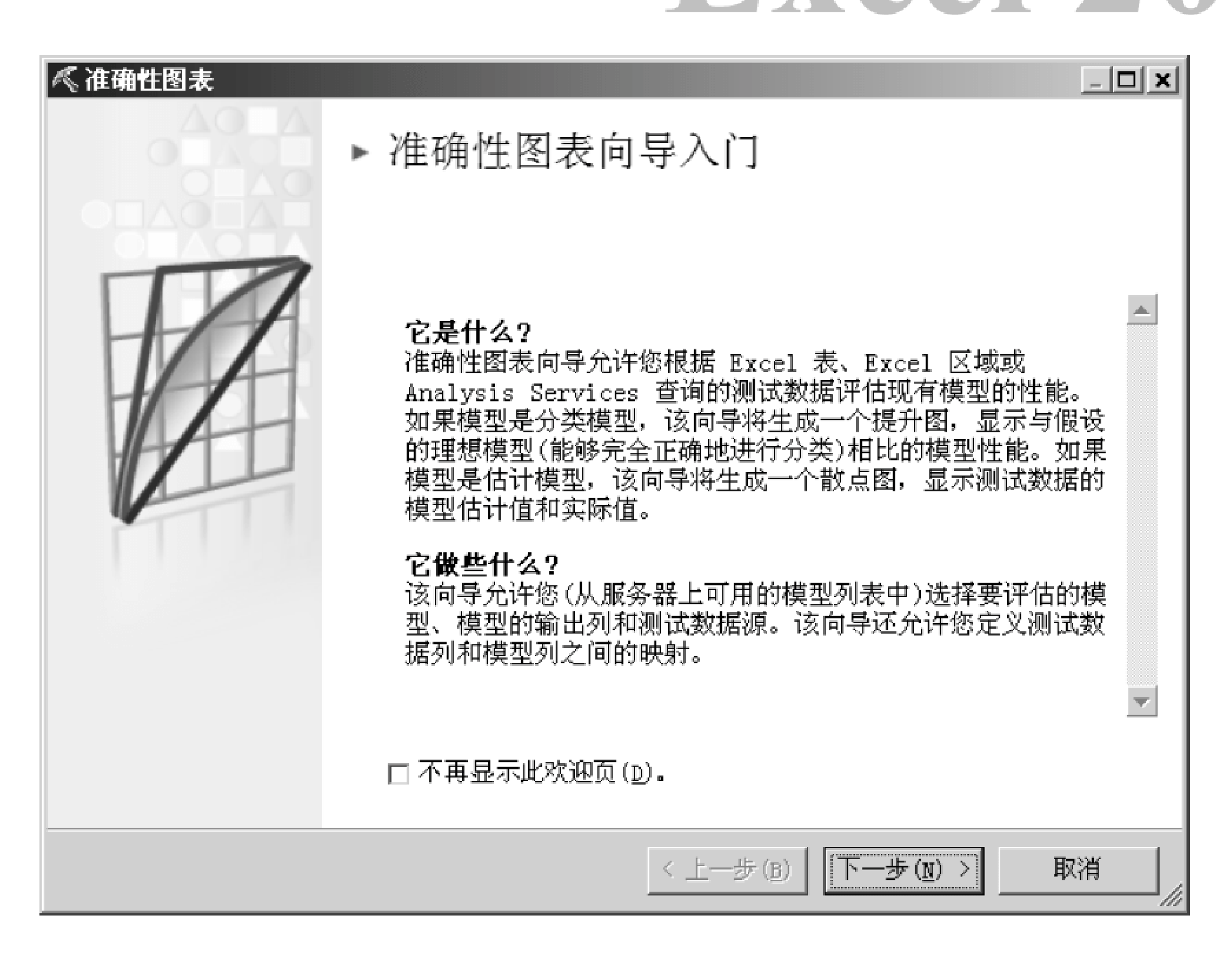


图 14-12 【准确性图表向导入门】窗口

Step9：在如图 14-13 所示的【指定要预测的列和要预测的值】窗口中，选择将要进行预测的挖掘列，本次选择 Purchased Bike 进入图表，单击【下一步】按钮。

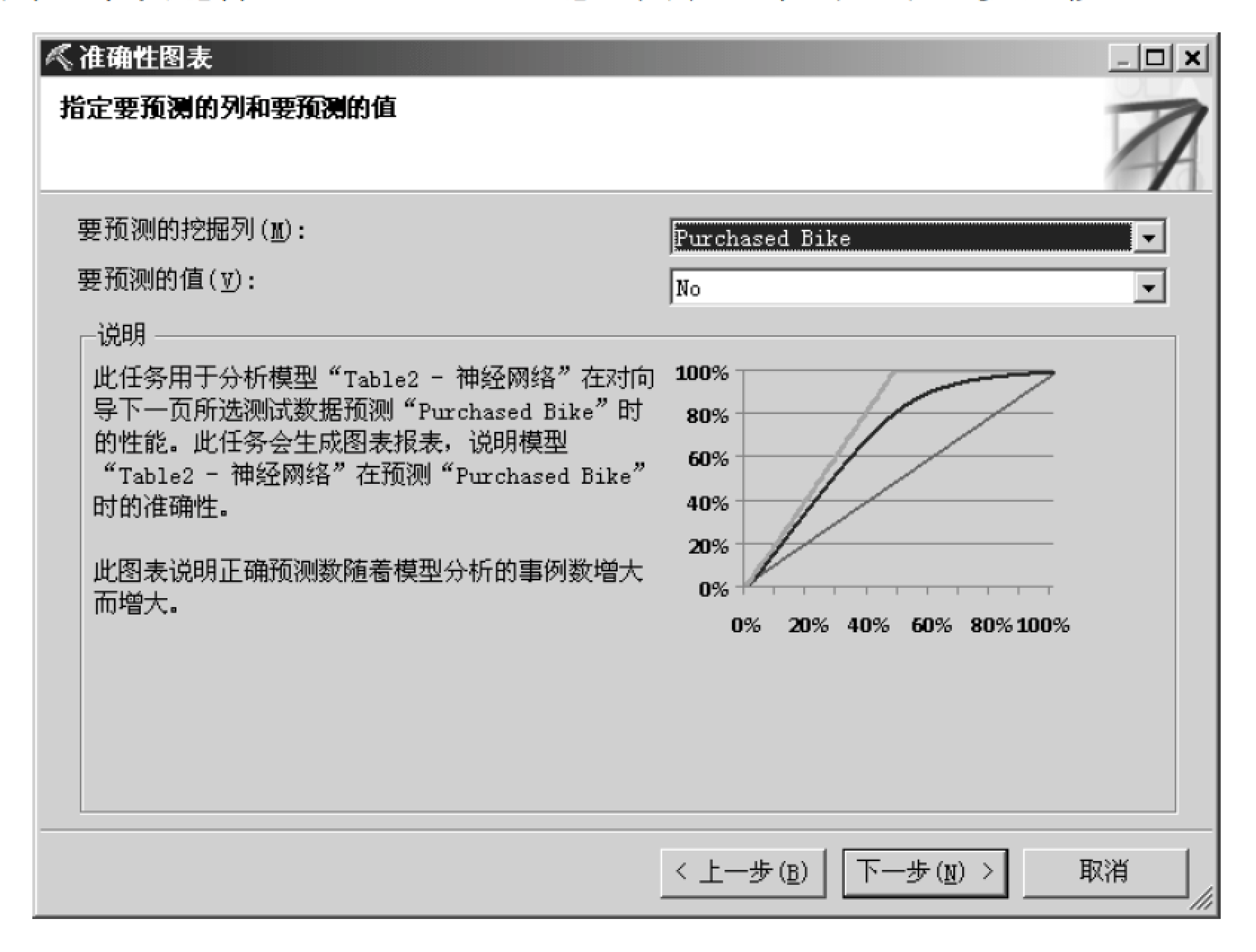


图 14-13 选择要预测的挖掘列

Step10：复制图表到 Excel 中，如图 14-14 所示。

Step11：单击【数据挖掘】中的【分类矩阵】按钮，弹出如图 14-15 所示的【分类矩阵向导入门】窗口，单击【下一步】按钮。

模型“Table2 - 神经网络”的准确性图表

预测列“Purchased Bike”= No

“Table2 - 神经网络”模型提升 121.89 %

图 14-14 复制到 Excel

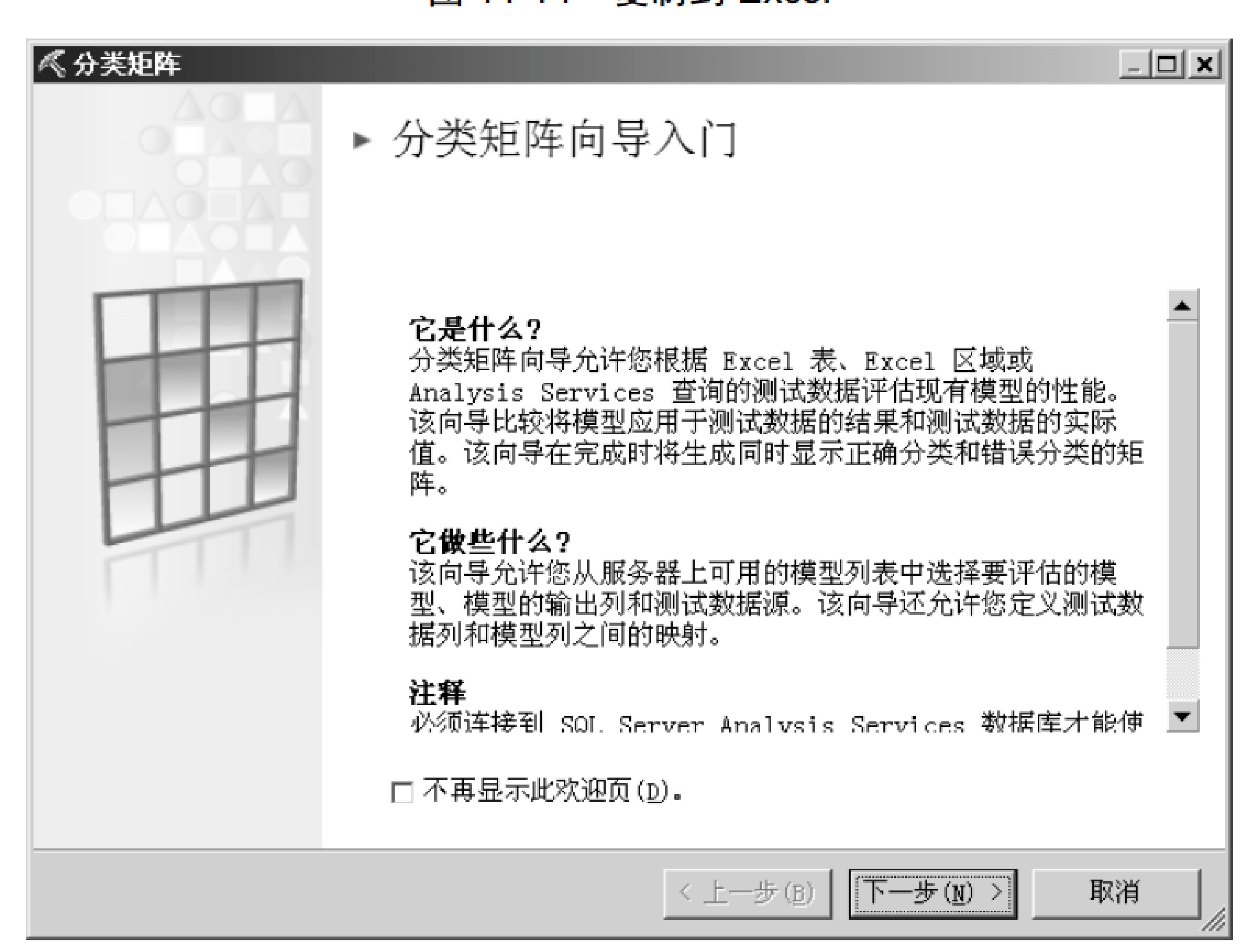


图 14-15 【分类矩阵向导入门】窗口

Step12：在【要预测的挖掘列】下拉列表框中，选择要预测的挖掘列，即自行车购买作为分析变量如图 14-16 所示，单击【下一步】按钮。

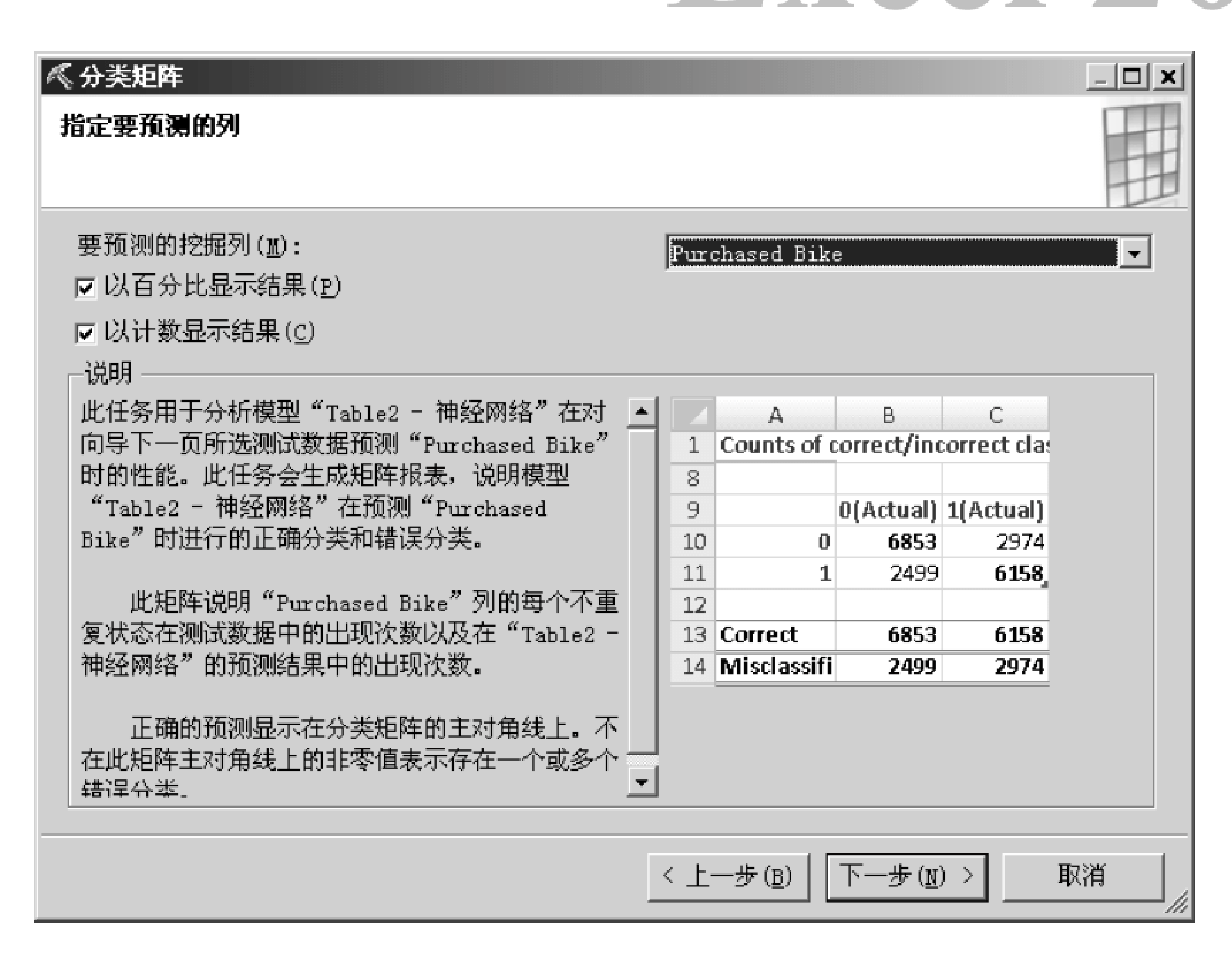


图 14-16　选择要预测的挖掘列

Step13：在如图 14-17 所示的【指定关系】窗口中，选择变量间关系，单击【完成】按钮。



图 14-17　【指定关系】窗口

Step14：产生分类矩阵，并复制到 Excel 中，如图 14-18 所示。

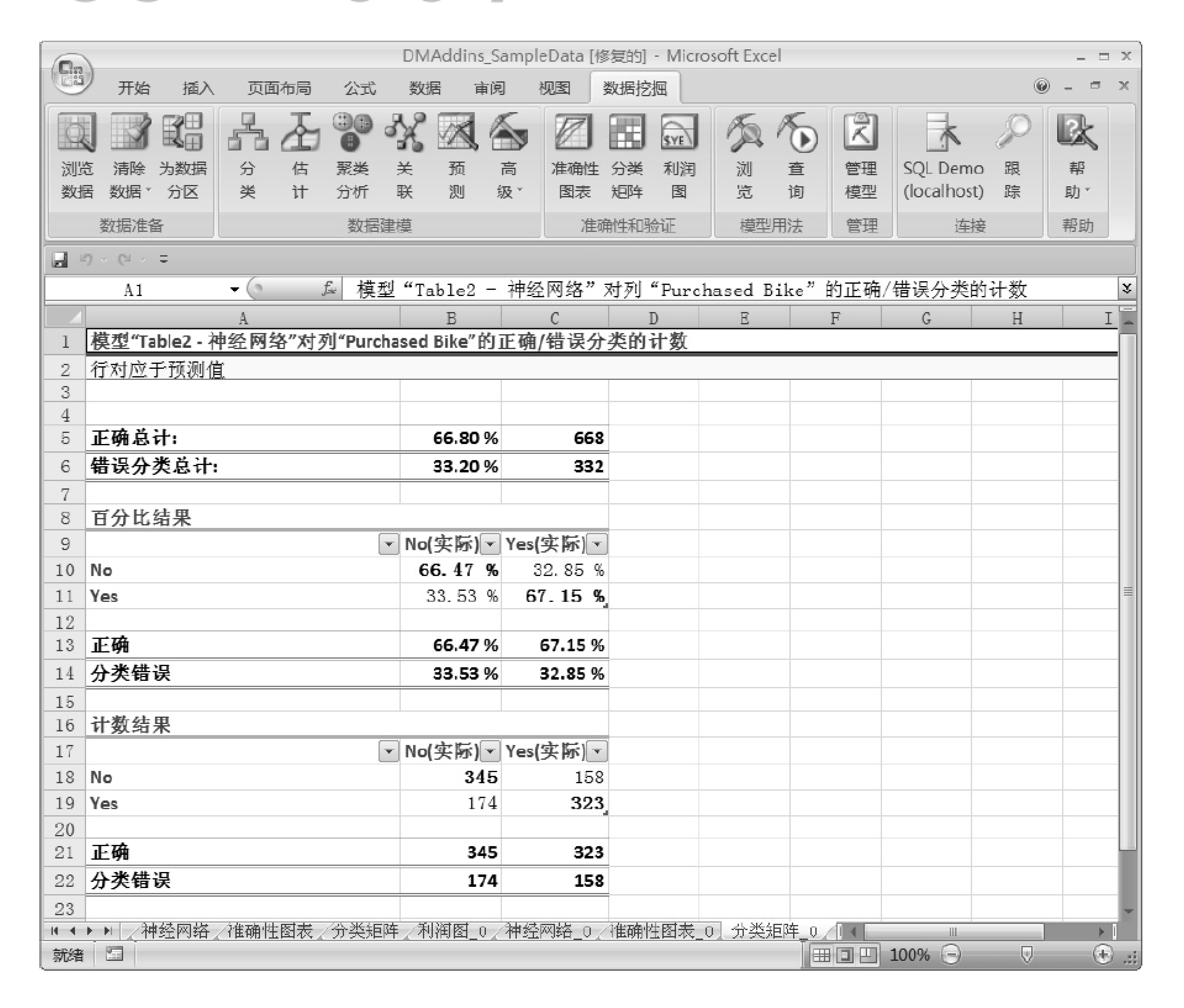


图 14-18 复制到 Excel

Step15：单击【数据挖掘】中的【利润图】按钮，弹出如图 14-19 所示的【利润图向导入门】窗口，单击【下一步】按钮。

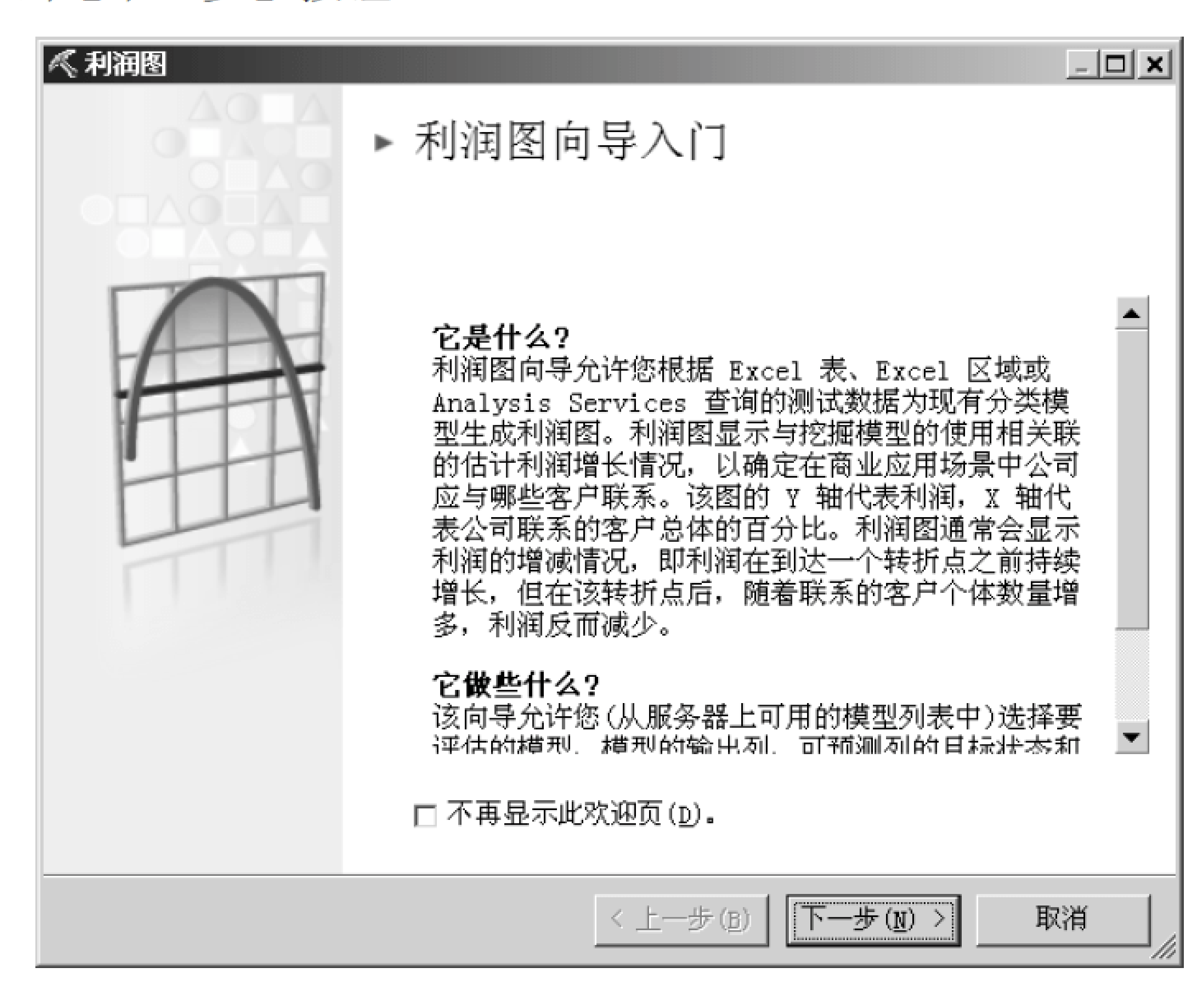


图 14-19 【利润图向导入门】窗口

Step16：在如图 14-20 所示的【指定利润图参数】窗口中，选择要预测的挖掘列，单击

【下一步】按钮。

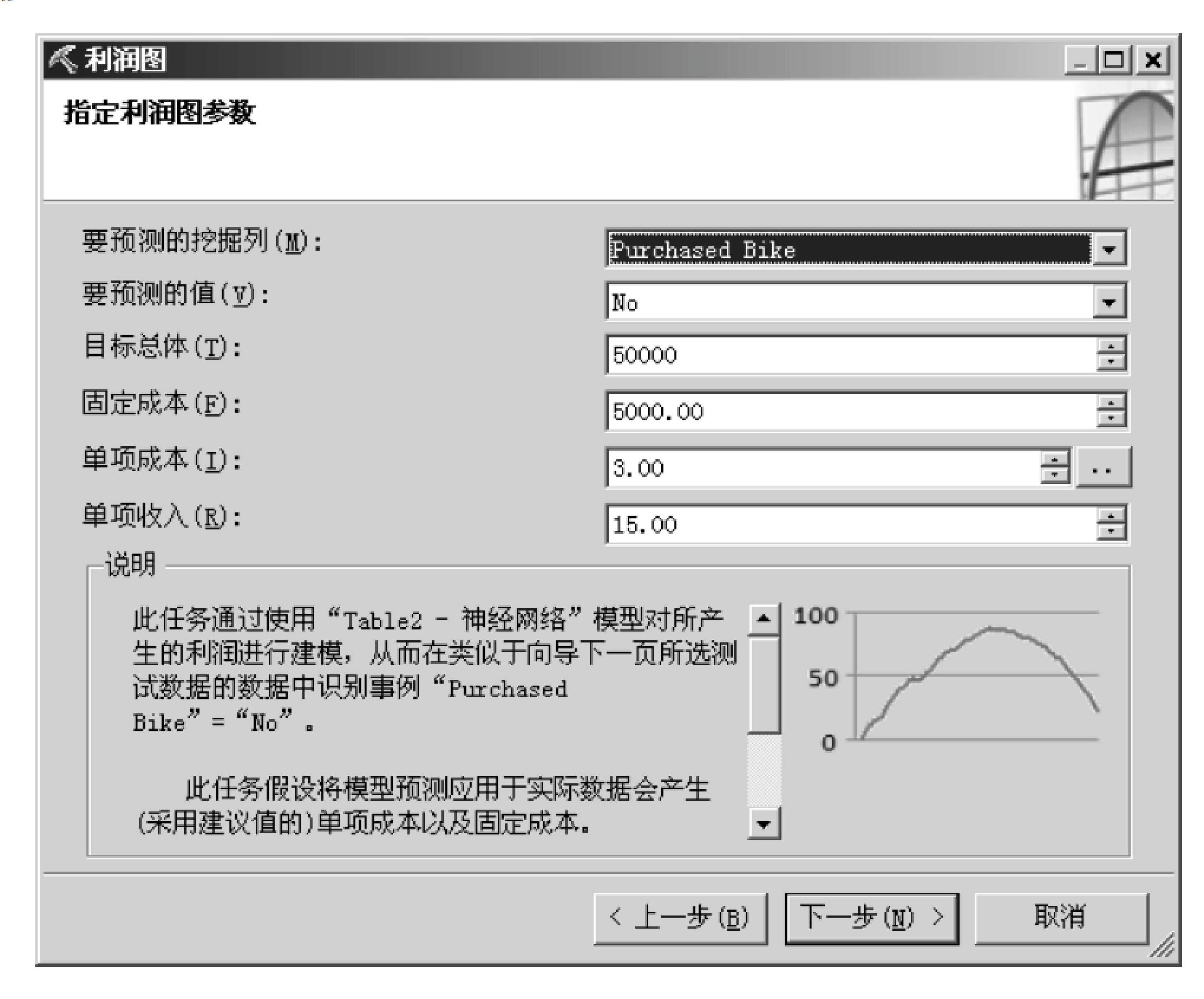


图 14-20 选择要预测的挖掘列

Step17：在如图 14-21 所示的【指定关系】窗口中，选择变量间关系，单击【完成】按钮。



图 14-21 选择变量间关系

Step18：将利润图复制到 Excel 中，如图 14-22 所示。

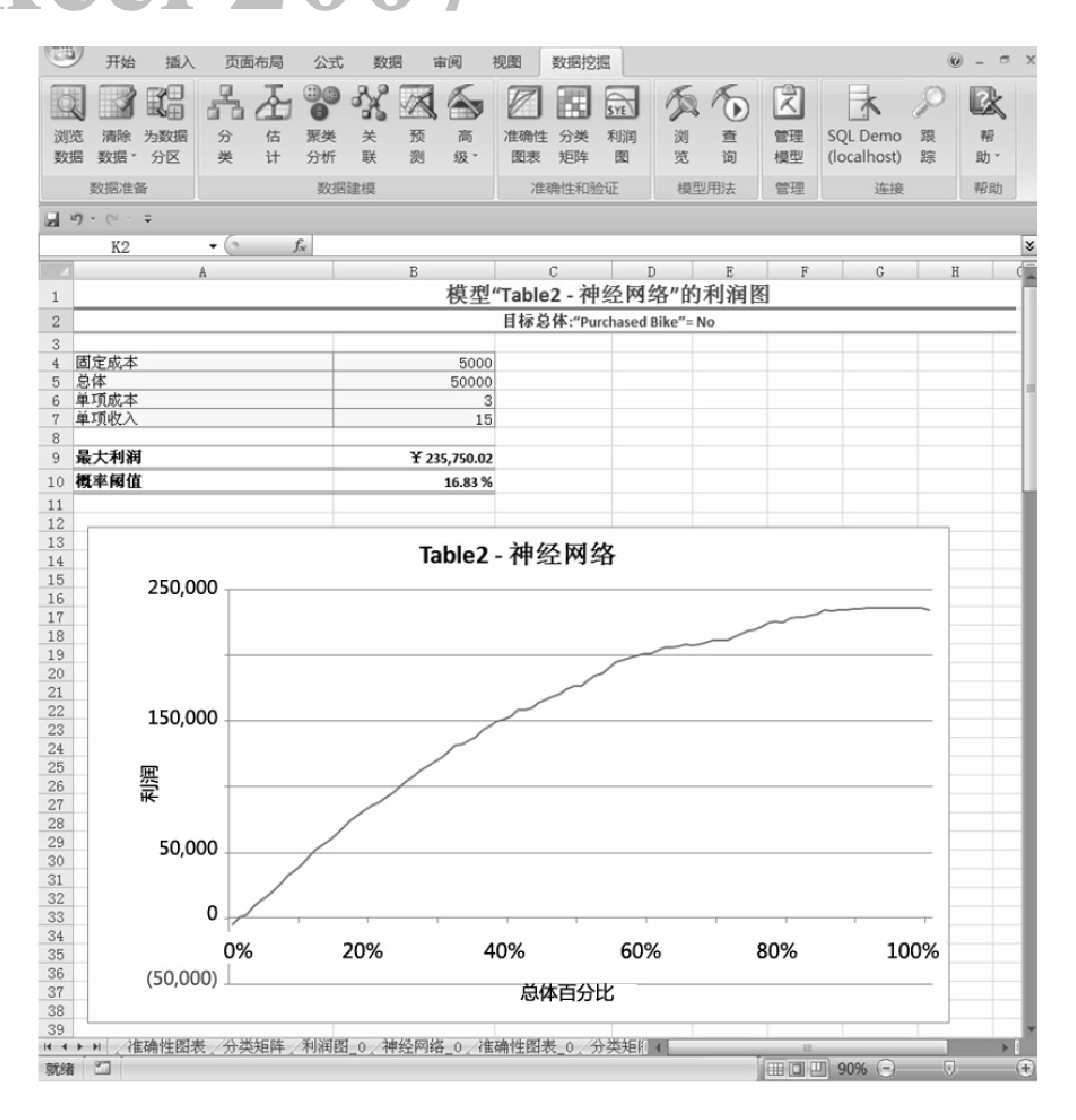


图 14-22　复制到 Excel

第 15 章　时间序列分析

15.1　基 本 概 念

观察值时常依时间的变化而发生一系列有规则的变化，这种数据称为时间序列数据，而对这种数据的分析方法称为时间序列分析法。人类社会的各种活动所产生的数据如果以发生的时间来区分，则可分为横截面数据（cross section data）及时间序列数据（time series data）两种。横截面数据是指发生于同一时期的数据；时间序列数据指的是同一个体的同一变量在不同时点或不同时期的数据，包括逐日的日数据、周数据、月数据、季数据及年数据等。例如：1990 年 10 月 3 日至 2006 年 8 月 4 日的上海 A 股指数数据。时间序列分析的目的在于观察、分析过去的数据，以预测未来。本章将介绍几种分析时间序列数据的方法。

预测方法可分为定量法与定性法两种。定量的预测方法是分析某个时间序列或可能与其相关的其他时间序列的历史数据的方法。若预测的方法仅限于使用该序列的历史数据值，则这种方法称为时间序列法。若在定量预测方法中所使用的历史数据涉及其他的时间序列，则应使用因果法。多元回归分析即为因果预测法。定性预测方法通常是运用专家的判断，这些程序的优点是可使用在无历史数据可供参考的情形，本书将在后面讨论这种程序。

时间序列分析已被各界广泛采用，其主要目的为：

① 对时间序列未来趋势作预测。

② 将时间序列分解成主要趋势成分（trend components），季节变化成分（seasonal component）。

③ 对理论性模型与数据进行拟合优度检验，以讨论模型是否能正确地表示所观测的现象，如一些常见的经济模型。

大部分时间序列分析法都先假设时间序列存在着某种数学结构，然后在此结构下延伸推导出分析结果来。一个时间序列常被假设为平稳型（stationary），或者是通过某些方法使其平稳，最常用的方法是差分法（differencing）。在探讨统计模型是否合适之前首要工作是先诊断时间序列的性质是否符合所使用方法的假设前提。然而，要检查一个时间序列是否符合时间序列分析的假设前提是一项艰难的工作，因此实证分析时经常以图形或以某些统计量对时间序列的基本性质做初步的判断。

在经济及商业方面，有许多应用时间序列分析法的实际例子，如国民生产总值（GNP）、失业率与股价等。而人们所关心的主题是去了解时间序列的行为，不仅是时间序列本身与过去的自我相关，还包括与其他时间序列的相关程度。这些时间序列最重要的共同特征是它们很少重复出现。一般可利用随机变量 x_1 构建时间序列 $x_1, x_2, x_3\cdots$，但是在时间序列的情况下这些变量 x_1, x_2, x_3，$\cdots$却仅能观测一次，这是与其他统计分析法不同的地方。

经济与商业时间序列的另一项难题是时间序列的结构常因政策变动或偶发事件而改

变。配合过去对时间序列的经验，有大量的文献探讨时间序列的理论观点。在 20 世纪 40 年代由 Norbort Wiener 和 Andrei Kolmogorov 提出平稳型时间序列的基本理论，而目前时间序列模型的研究也已转向较具有应用性的课题方面，对此项转变有重要贡献的学者有 Whittle、Quenuille、Rosenblatt、Parzen、Hannan、Box、Grenander、Rozanov、Granger、Tiao 等。

在 20 世纪 60、70 年代，一份工程文献提出了新的时间序列的技巧，文献的作者是 Kalman Kailath、Lennart Ljung 和 B.D.O. Anderson。他们所强调的时间序列分析法与统计学家及经济学家的略有不同。由于工程学的研究数据经常是庞大的，所以他们对于过滤法（filting）、平滑法（smoothing）及计算法（algorithm）的发展很有兴趣。另一方面，统计学家则花了许多心思在模型的构建上，参数的估计和数据的拟合优度检验，其在推导的过程中仅需要适中的观测值个数，而不像工程学那样庞大。

从此应用时间序列分析法产生了两个分支，第一种是着重于时间序列的谱密度(spectral density）及频域分解（frequency domain decomposition）的频域法（frequency domain approach)，这是一门运用非参数统计的时间序列分析方法，常应用于自然科学方面，如工程学和物理学，但在经济学方面也开始受到重视。由频率定义分析所得的结果常被视为系统中基本的变动。

第二种时间序列分析法则是利用时间序列的参数模型（parametric modeling）的 arima（autoregressive integrated moving average）模型及较为复杂的多变量 arma 模型，而 arma 模型则包含两个重要的子模型 ar（autoregressive）和 ma（moving average）模型。

当利用 ARMA 模型对一平稳型时间序列建模时，即是利用其参数的结构来描述数据的记忆型态。此法则能在建模时仅需利用有限个参数，相较于非参数的光谱密度法来说可使参数的估计更合理可行，且需要的观测值个数也较少。而利用参数建模时更提供了一种由历史数据预测时间序列未来趋势的实用方法。

此外，可利用差分及过滤法对非平稳型（nonstationary）时间序列建模。在时间序列建模时，最重要的观念是如何利用过去的数据来判定一个变量的未来走向及不同变量间的同期（concurrent）或前后期（lead-lag）的关系。

相较于过去传统的 Box 和 Jenkins 单变量时间序列模型，近来已有许多学者对多变量时间序列模型进行研究，例如 Box 和 Tiao （1982）及 Tiao 和 Tsay（1983）。

多变量时间分析法的研究含有两种目的：一是加入另一个相关的时间序列后，更能解释过去仅由单变量建模的不足之处；另一个目的则是通过分析一个时间序列与另一个时间序列的关系，借以获得时间序列间的相关信息，来增进对整体系统的了解。

近 15 年来在非线性及多变量时间序列分析法的领域中有许多新的进展，较为重要的研究课题包括 ARCH Models、Threshold AR Model、Co-Integration、Reduced Rank Models、Scalar Component Models 和 State-Space Models。在本书中引用了 Box 在 1980 年提出的高级建模技术并且探究以递归方式对时间序列数据构建模型。

时间序列具有如下几个特性：

① 时间序列中的观测值由四个影响成分组成，分别是长期趋势（trend）、循环变动（cyclical fluctuation）、季节变动（seasonal fluctuation）、不规则变动（irregular fluctuation）。

因此进行时间序列分析时应先将这四个成分分解出来，以了解各个成分的影响。

② 时间序列的各个观测值通常互有关联，只是时间相隔越长，关联越小。

③ 因分析需要，不同时间单位的时间序列数据，可以转换成相同时间单位的时间序列。例如，年数据转换为月平均数据。

④ 时间序列应依时间先后顺序排列，不可任意变更。

⑤ 时间序列的时间单位可以为年、季、月、周、日等，应划分为相同间隔的时间单位。

时间序列的数据在分析前，须将数据按时间次序，以纵轴为变量，横轴为时间作图，此图称为时间序列图，如图 15-1 所示。从此图中可大致看出时间序列的特性，即使相似的频数分布图，时间序列的变动也可能不同，如 *A*、*B* 两时间序列虽有相似的频数分布，但其时间序列的变化并不相同。

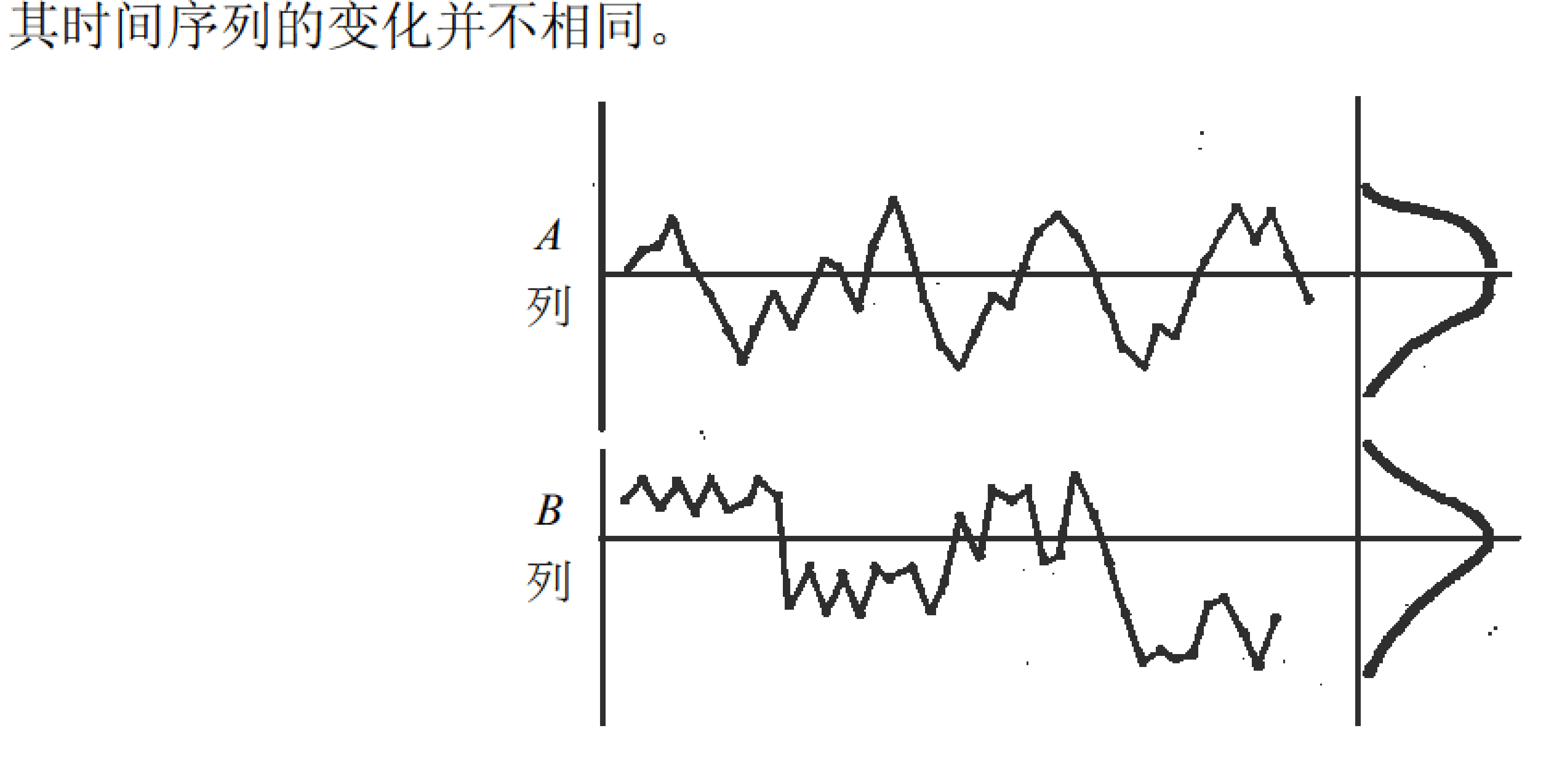


图 15-1　时间序列与频数分布图

15.2　时间序列的成分

通常时间序列是由四个成分——趋势、循环、季节与不规则组成。

1．趋势成分（trend component）

时间序列分析的测量数据，可取自于每一小时、天、星期、月或年，或任何其他有规则的区间，限制序列的记录值是来自相等的区间，若是不相等区间的观察值的处理问题，则超出了本书的范围。虽然一般的时间序列数据显示随机的上下变动，但就长期来看，它仍然逐渐地增高、降低或在一定范围内变动，这种逐渐变动的时间序列，经常是由于长期因素所导致的，例如人口的变动、人口统计上的特征改变、工业技术的改进等，称之为时间序列的趋势。

2．循环成分（cyclical component）

当时间序列在长期间里显示某种趋势时，不能预期所有时间序列的未来值将落在趋势线上。事实上时间序列的观测值经常落在趋势线上方与下方。落在趋势线的上方与下方的序列点的任何周期超过一期的有规则的模型皆属于时间序列的循环成分。

许多时间序列的连续观察值规则地落于趋势线的上方与下方，而显示循环的现象。一般相信在经济上多期的循环变动，可以用这种时间序列的成分来代表。

3．季节成分（seasonal component）

虽然时间序列趋势与循环成分往往通过分析多年的历史数据才能辨认，但有许多时间序列在一年内即显示周期性的规则变动。例如，游泳池的制造商可以预测其在秋冬季的月份中，销售收入较低，在春夏季的月份则销售收入较高。这种随着季节的影响而变动的时间序列成分，称为季节成分。一般都认为时间序列的季节变动是在一年之内，然而常用它来分析一年之内的连续重复数据。例如每天的交通流量也显示了一天内的“季节”情况，在尖峰时间为最拥挤，白天的其他时间及傍晚流量为中等，而从午夜至凌晨则流量为最低。

4．不规则成分（anomial component）

时间序列的不规则成分就是当完全以趋势、循环及季节等分量来说明此时间序列时，用来解释实际的时间序列值与所预期的序列值之间的离差的残差因素，它是用来说明时间序列的随机变动。时间序列的不规则成分，常是由短期不可预知或非重复的因素所引起的，它是用来说明时间序列的随机变动，所以无法预测，更无法预知它对该时间序列的冲击。

时间序列的四个组成分子的关系可分为两种模型。

（1）相加模型（additive model）：$Y=T+S+C+I$

① 模型中所有的数值均以原始单位表示。

② 若 $S>0$ 表示季节变动对 Y 有正的影响。

③ 若 $C>0$ 表示景气循环正在衰退。

④ 若 $I>0$ 表示有些随机事件对 Y 有正的影响。

相加模型的最大缺点是假设各个组成部分彼此独立，然而现实生活中，任一个部分变动有时会影响其他部分的变动，因此在经济活动中，此模型并不适合。

（2）相乘模型（multiple model）：$Y=T\times S\times C\times I$

① 模型中 T 以原始单位表示，C、S、I 以百分比表示。

② C、S、I 均大于 1 时表示相对效果高于趋势值，若小于 1 时表示相对效果低于趋势值。

③ 相乘模型假设各个组成部分彼此相互影响，非独立。

④ 由于季节变动只发生于一年，因此对于年数据的相乘模型为 $Y=T\times C\times I$。

15.3 时间序列数据的图形介绍

图 15-2～图 15-7 表示一些时间序列数据的图形。图 15-2 为连续观测一项化学反应的 70 笔产量的观测值，这 70 笔的时间序列数据的明显特征就是在一固定的水平为 50 左右，并且在 20～80 的固定限度内变动，大致上序列不论何时都维持相同的行为，除了在实验过程中发生基本的改变之外，对此类时间序列的预测可以序列的平均值为准。在此例中，所预测的产量的平均水平应为 50，且都在 20～80 之间。若再仔细观察序列的行为会发现一

个趋势：若观测值大于平均数，则下一个观测值即小于平均数，反之亦然，于是两两邻近的观测显示负相关，如果适当利用此相关性可使预测更精确。

例如：最后一个观测值小于平均水平，于是可预测下一个观测值应大于平均水平，而再下一个观测值应小于平均水平，如此循环下去，只要能够找到一个合适的概率模型（probabilistic model）来描述观测值在时间上的相依性，必然能使预测值更精确。然而如图 15-2 所示的平稳型序列在商业应用领域很少出现，较常遇到的数据类型是如图 15-3～图 15-5 这样的数据。

图 15-3 是每个月的电冰箱需求量，图 15-4 为美国 1800—1981 年每年的利率及物价指数，相比于图 15-2 的化学反应的实验数据，这些时间序列表现了一种随机游走的行为，此种时间序列称为无定向型序列或非平稳型序列（non-stationary series）。由于此种时间序列的平均水平本身随时间改变，因此无法再以一个固定的值来预测未来的变动。

此种时间序列的模型不同于图 15-2 时间序列的模型，当然预测的方法也有所不同。图 15-5 是由美国联邦储备委员会（Federal Reserve Board）出版的美国月度工业生产指数，由图发现该时间序列的行为有持续上升的趋势，所以可拟合出一条直线来拟合数据。然而若仔细观察数据走势可画出三条平行直线，第一条表现的区间为 1947—1960 年，第二条为 1961—1975 年，第三条则为 1975—1993 年。所以如何找出一种合适的概率模型来拟合这些平行线，并且由模型如何去预测未来的数值是需要探讨的。

最后一个例子，图 15-6 是每月国际航线旅客总数取对数得到的数据，图 15-7 是 Magnavox 彩色电视每月销售量的数据，这些数据最明显的特征是具有季节性的变化行为，而大部分的原始商业数据常见这种季节性变化的行为，季节性的行为清楚地表现出每隔 12 个月数据的相依性，因此在构建合适的预测模型时，不仅要考虑每个月间的相关性，更需考虑同一月份在不同年之间的相关性。

上述的各种例子说明不同类型的时间序列数据需要创建不同类型的模型，并且没有任何一个预测模型能够适合所有的时间序列，所以在上述的例子中，所需做的就是建立一个能够合适地表达数据时间相依关系的概率模型，一旦建立此概率模型后，便可做有效的预测了。

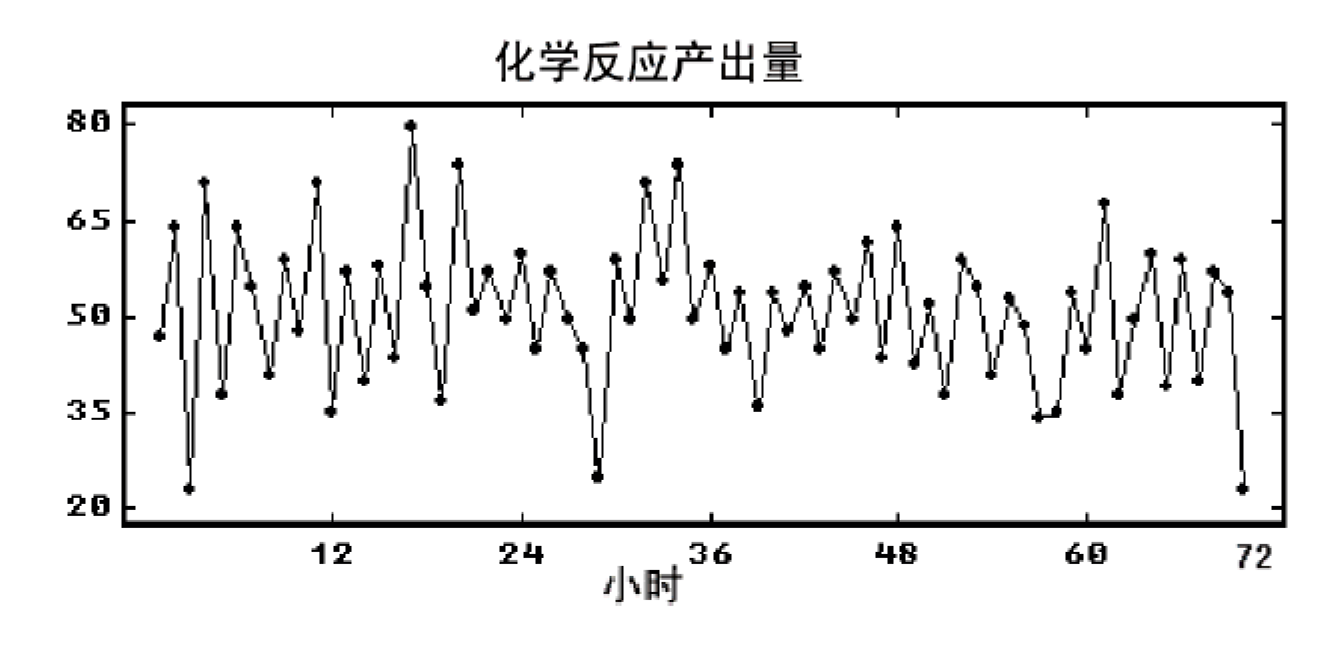


图 15-2　化学反应产出量（每次观测间隔两小时）

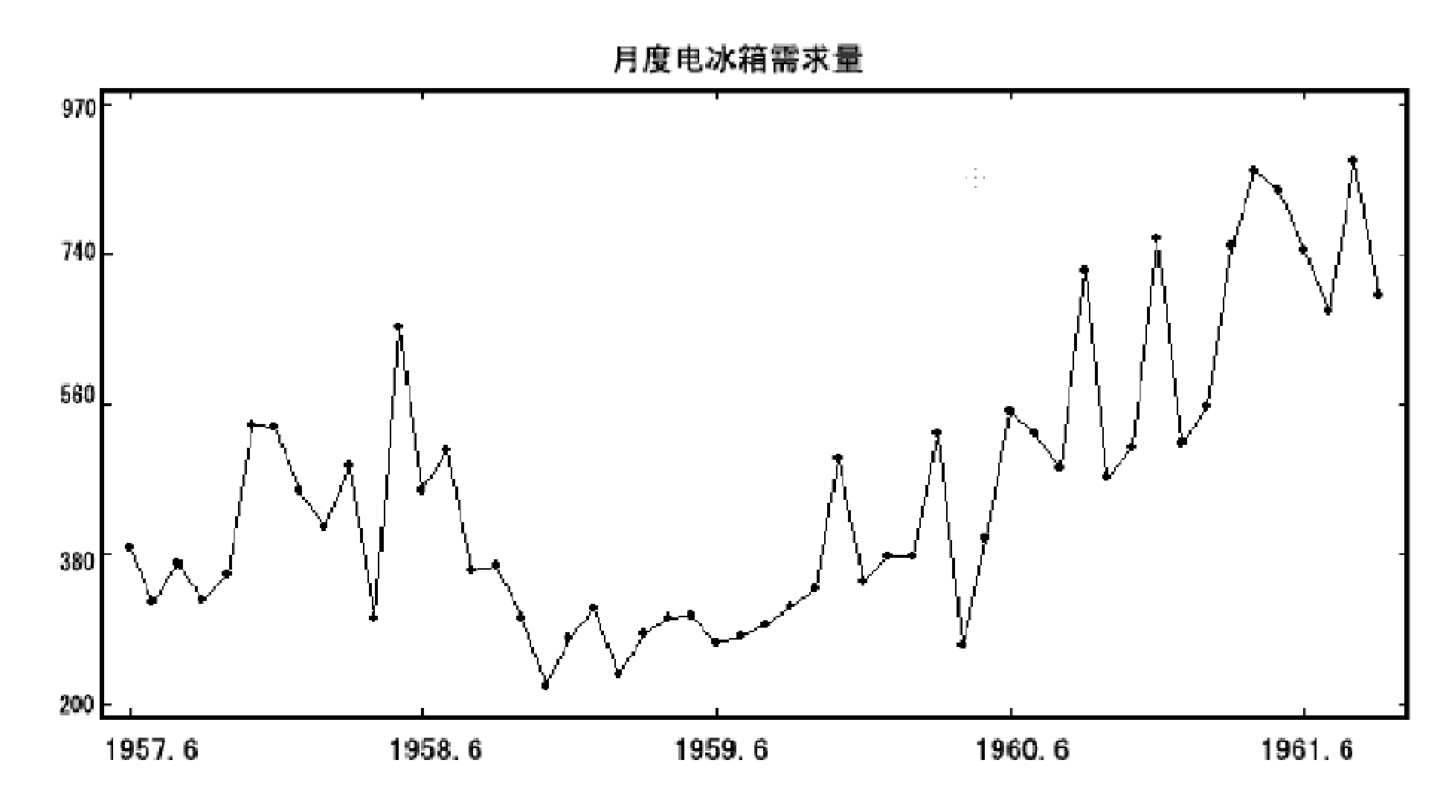
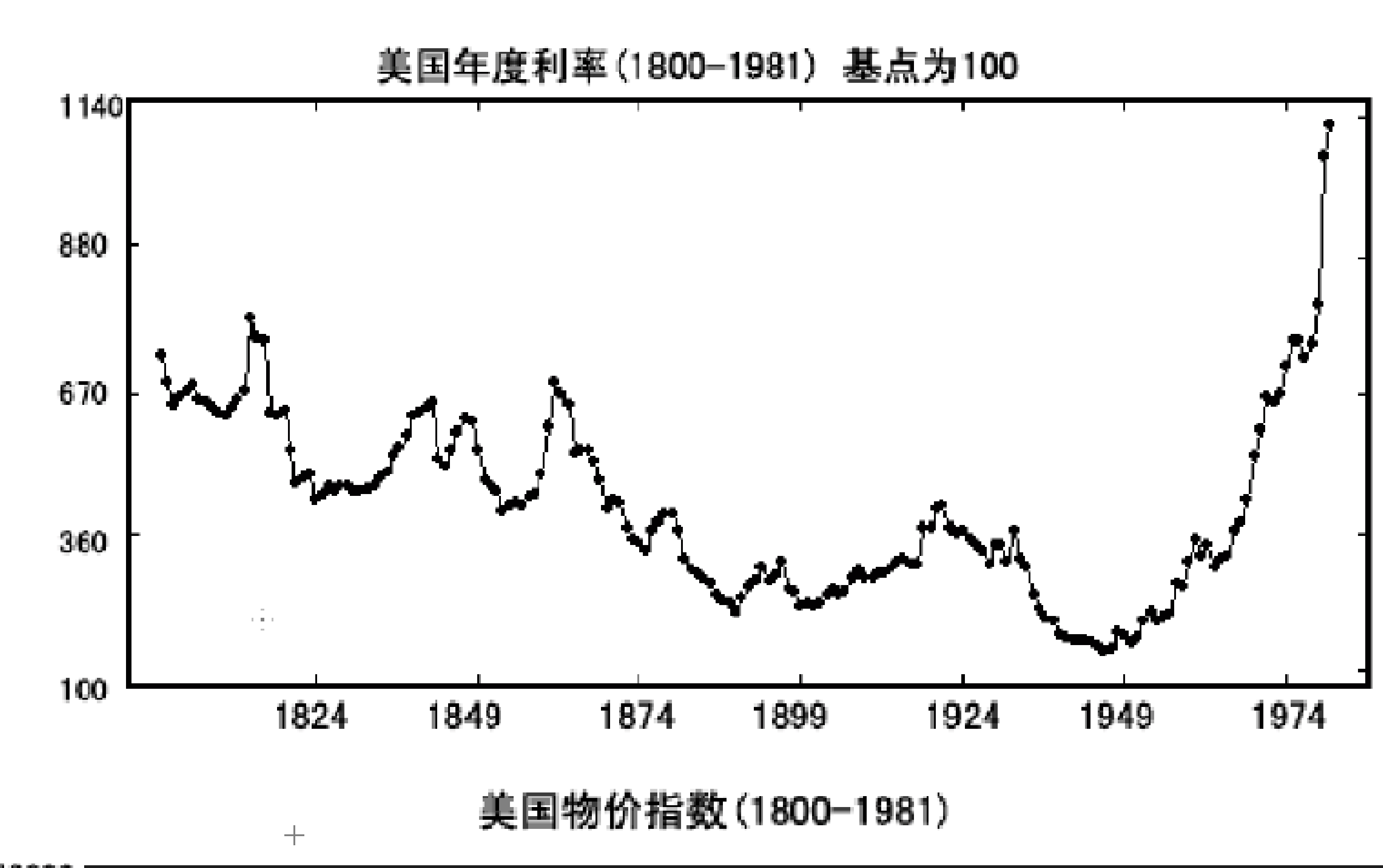


图 15-3　美国电冰箱月度需求（千台）（1957 年 6 月—1961 年 9 月）



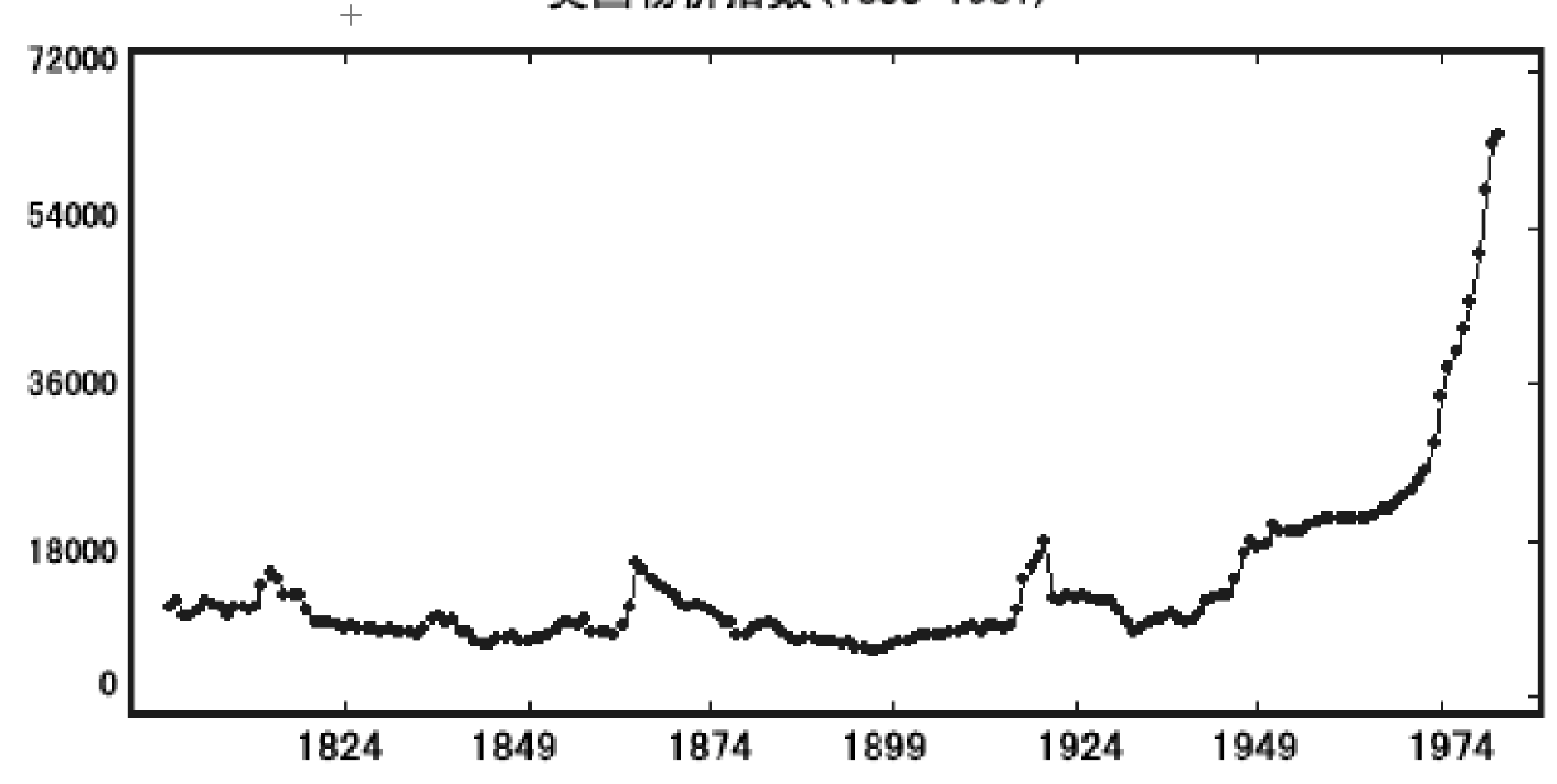


图 15-4　美国年度利率与物价指数

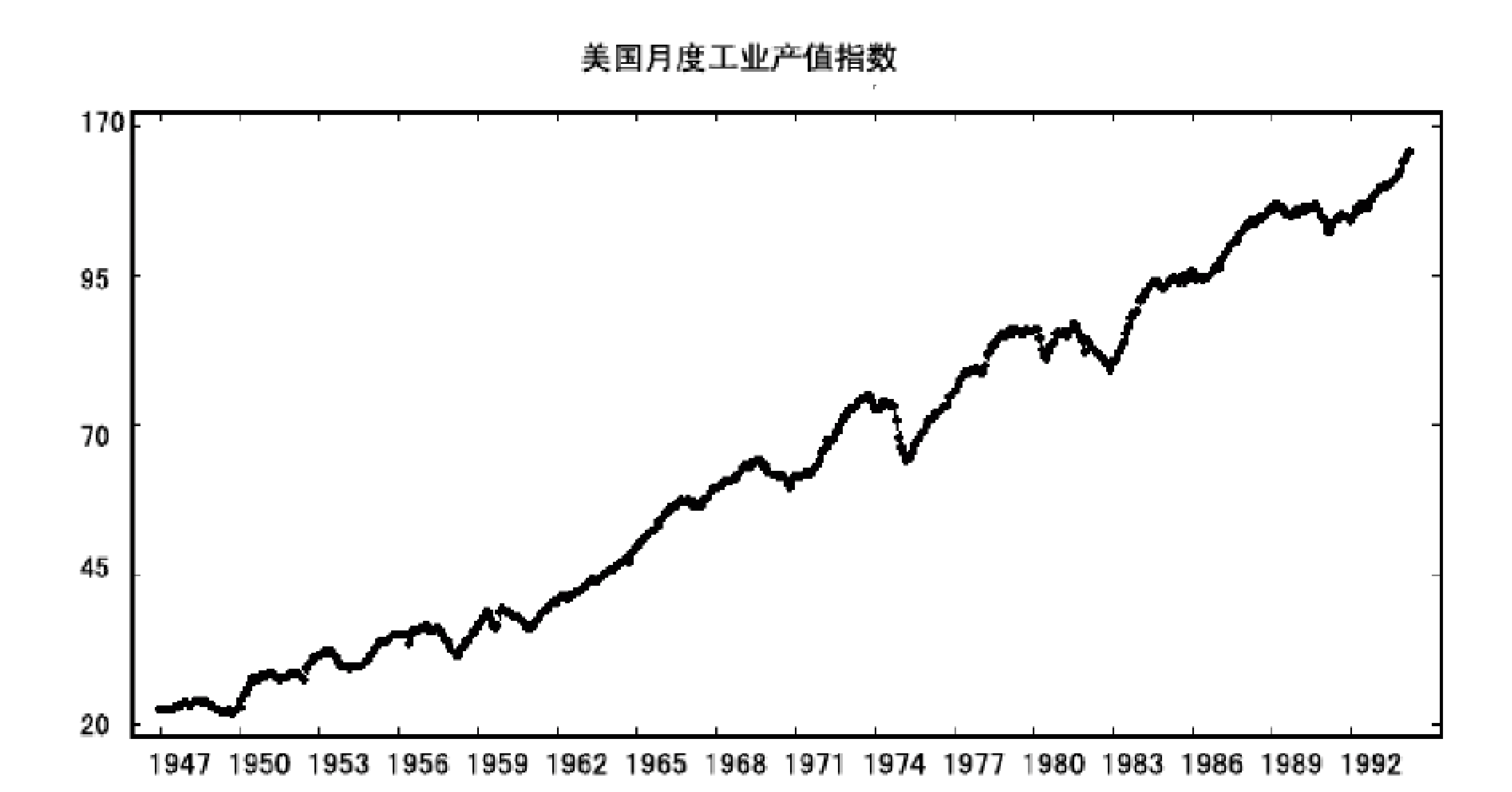
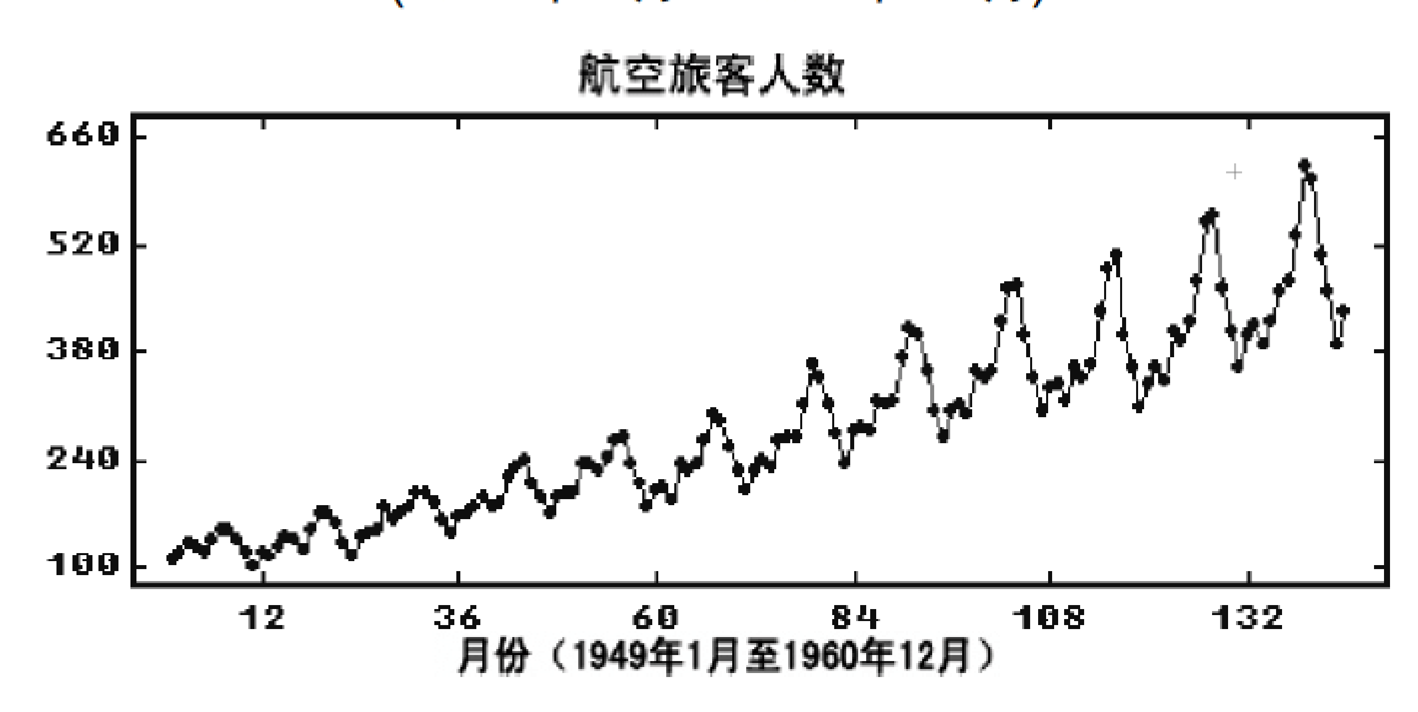
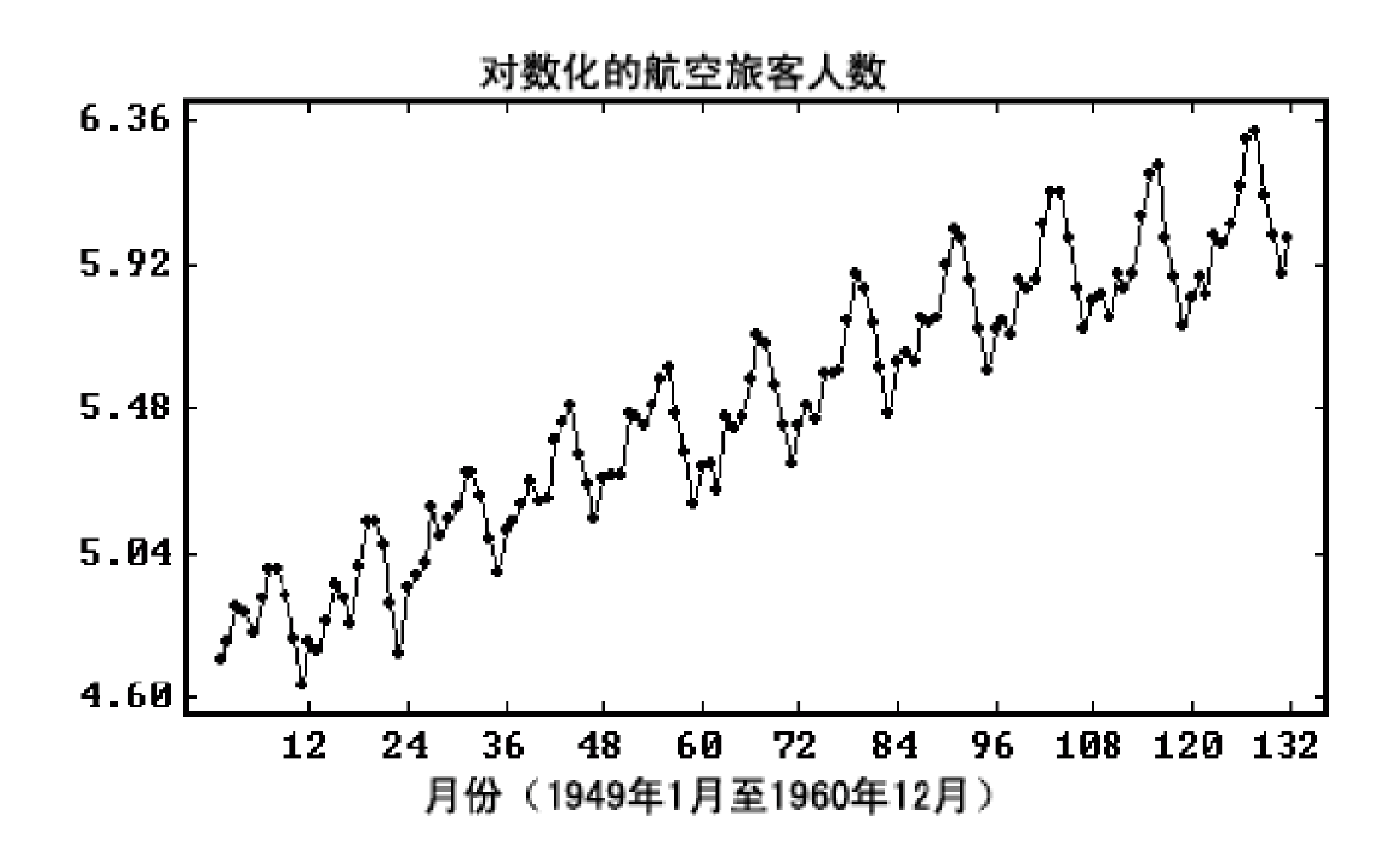


图 15-5 美国月度工业产值指数

(1947 年 1 月—1993 年 12 月)



对数转换后航空旅游人数



*取自 Box-Jenkins(1976)序列 G。

图 15-6 美国月度国际航空旅游人数

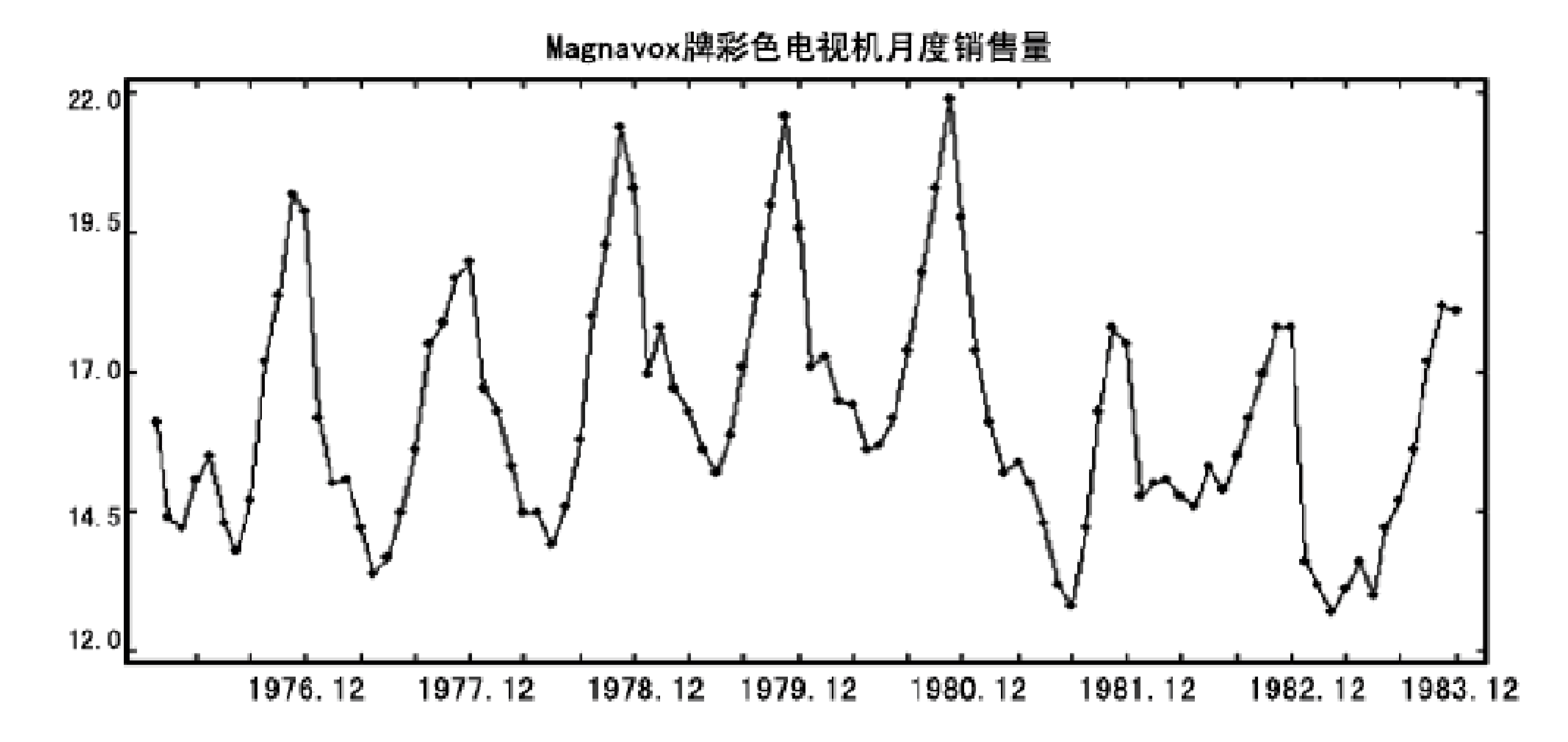


图 15-7　Magnavox 牌彩色电视机月度销售量（千台）（1976 年 1 月—1983 年 12 月）

总结时间序列的形态大致包含下述五种：

① 平稳型（stationary）。

② 无定向型（drifting）。

③ 趋势型（trend）。

④ 季节型（seasonality）。

⑤ 外部影响型（exogenous effect）。

15.4　利用平滑法预测

本节将讨论适于无明显趋势、循环或季节效应的时间序列的预测方法。在这种情况下，预测方法的目的是以平均过程“修匀”时间序列的不规则成分。首先考虑移动平均。

1．移动平均

移动平均法是将最近 K 期的时间序列数据加以平均，以所得平均值预测下一期的数据。此种预测方法很简单，适用于不明显长期趋势与季节循环变动的时间序列数据。

移动平均的计算公式如下：

$$x_t = \frac{1}{n^0}\sum_{i=1}^{n^0} x_{t-1}$$

然而此种方法的准确性，则必须用预测值与观测值之间的误差来衡量。一般用来衡量预测误差大小的公式为平均平方差（mean square error，MSE）：

$$\text{MSE} = \frac{1}{n^0}\sum_{t=1}^{n^0}(Y_t - \hat{Y}_t)^2$$

式中 Y_t 为真实观察值，$\hat{Y}_t$ 为预测值，n^0 为预测的期数。

期数 n^0 的选择影响预测的精确度，一般而言，大都采用“尝试错误”的方法，即采取 $T_1, T_2, T_3,\cdots$等期计算移动平均，并计算 MSE，比较 MSE 的大小，选择具有较小的 MSE 的期数。

例如某中介公司的 12 周租赁中介量如表 15-1 所示，预测第 13 周销售量。

表 15-1　12 周租赁中介量

周次	1	2	3	4	5	6	7	8	9	10	11	12
销售量	63	81	72	63	54	72	87	84	60	48	60	66

由表 15-1 可知无明显长期趋势及季节变动，所以利用移动平均法来预测销售量。

首先假设期数 $K = 3$（以 3 周数据计算移动平均）：

以第 1～3 周销售量平均预测第 4 周，(63+81+72)/3=72。

以第 2～4 周销售量平均预测第 5 周，(81+72+63)/3=72。

逐一计算可得表 15-2 第 3 列数据。

计算误差大小如表 15-2 第 4、5 列数据，由此可知：

$$\text{MSE} = \frac{1}{n^0}\sum_{t=1}^{n^0}(Y_t - \hat{Y}_t)^2 = 2\,629/9 = 292.1$$

表 15-2　中介公司 3 周移动平均预测值

周　　次	时间序列值 Y_t	移动平均预测值 $\hat{Y}_t$	预测误差 $Y_t-\hat{Y}_t$	预测误差平方 $(Y_t-\hat{Y}_t)^2$
1	63			
2	81			
3	72			
4	63	72	–9	81
5	54	72	–18	324
6	72	63	9	81
7	87	63	24	576
8	84	71	13	169
9	60	81	–21	441
10	48	77	–29	841
11	60	64	–4	16
12	66	56	10	100
13		58		$\sum$=2 629

2．加权移动平均

在移动平均法的计算中，每一个观察值均具有相同的权数。另一种可能的方法，即熟知的加权移动平均，它是赋予每个数据值不同的权数，而后再以加权平均作为预测值。加权移动平均法是依据各期的重要性，给予不同的权数（weight）用以计算 K 期移动平均数。

当要预测某一期的数值时，通常最近一期的影响最大，而前几期的影响较小，因此最近一期的权数应与其他前期的权数不同。在大部分情况中，将最大的权数放在最近的观测值上，且权数随着数据值的久远而递减。

在此要注意的是，加权移动平均的权数总和要等于 1，对简易的移动平均而言，也是如此。

上例加权移动平均法计算结果如表 15-3 所示。

表 15-3　中介公司 3 周加权移动平均预测值

周　　次	时间序列值 Y_t	移动平均预测值 $\hat{Y}_t$	预测误差 $Y_t-\hat{Y}_t$	预测误差平方$(Y_t-\hat{Y}_t)^2$
1	63			
2	81			
3	72			
4	63	73.5	–10.5	110.25
5	54	69	–15	225
6	72	60	12	144
7	87	64.5	22.5	506.25
8	84	76.5	7.5	56.25
9	60	83	–23	529
10	48	72.5	–24.5	600.25
11	60	58	2	4
12	66	56	10	100
13		61		$\sum$=2 275

3．指数平滑法

指数平滑法是利用过去时间序列的加权平均值来平滑数据的方法，并利用该加权平均值作为下一期的预测值。

以下仅介绍最简单且常用的一阶加权平均数（first order exponential smoothing method），公式如下：

指数平滑模型：$F_{t+1}=\alpha Y_t+(1-\alpha)F_t$

其中：

F_{t+1}：t+1 期的时间序列预测值。

Y_t：t 期的时间序列实际值。

F_t ：t 期的时间序列预测值。

α：加权系数($0\leqslant\alpha\leqslant1$)。

若时间序列的随机变异较大时，则加权系数α应较小，以避免因大的预测误差而影响预测值。

F_1为第 1 期的预测值，但因无第 1 期以前的数据，故一般以第 1 期的观测值为预测值。

距预测期越近的观测值对预测值影响越大，距预测期越远的观测值对预测值影响越小，因 $Y_t, Y_{t-1},\cdots Y_1$ 的权数随着$(1-\alpha)$的指数增加而递减。较小的 α 值产生较平滑的曲线，较大 α 的值产生较不平滑的曲线。

为让读者了解任意一期的预测值为所有前期的实际值的加权平均，假设有一含三期数据 Y_1、Y_2、Y_3 的时间序列。首先令 F_1 等于第 1 期的时间序列实际值，也就是说，$F_1=Y_1$。因此，第 2 期的预测值如下所示：

$$\begin{aligned} F_2 &= \alpha Y_1 + (1-\alpha)F_1 \\ &= \alpha Y_1 + (1-\alpha)Y_1 \\ &= Y_1 \end{aligned}$$

一般而言，第 2 期的指数平滑预测值等于第 1 期的时间序列实际值。要得到第 3 期的预测值，将 $F_2=Y_1$ 代入 F_3 的表示式中，则得：

$$F_3 = \alpha Y_2 + (1-\alpha)Y_1$$

最后，将此 F_3 的表示式代入 F_4 的表示式中，得到：

$$\begin{aligned} F_4 &= \alpha Y_3 + (1-\alpha)[\alpha Y_2 + (1-\alpha)Y_1] \\ &= \alpha Y_3 + \alpha(1-\alpha)Y_2 + (1-\alpha)^2 Y_1 \end{aligned}$$

由此可以获知 F_4 是前三个时间序列值的加权平均，并且注意到 Y_1、Y_2、Y_3 的系数或权数的和为 1。对于任一个 F_{t+1} 的预测值而言，一样可以依此推导出，它是前面 t 个时间序列值的加权平均。

指数平滑法的好处是仅需要极少的历史数据来做简易的处理，只要加权系数 α 一经选定，则要计算下一期的预测值，只需两项数据。如同式中，若 α 已给定而要求 t+1 期的预测值，只要知道 t 期的实际值 Y_t 与预测值 F_t 即可。

事实上只要 α 值介于 0 与 1 之间皆可以。当然有某一些值比其他的值更好，而从下面改写的基本指数平滑模型中，可以选择出一个良好的 α 值：

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t$$

$$F_{t+1} = \alpha Y_t + F_t - \alpha F_t$$

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

由此知道新的预测值 F_{t+1} 等于前期的预测值 F_t 加上一项修正项即是 α 倍的最近预测误差 Y_t-F_t，也就是第 t+1 期的预测值是由第 t 期的预测值经预测误差修正而获得的。若这一时间序列的随机变异很大，则加权系数应选较小的，其理由是因为预测误差大部分是由随机变异引起的，因此不希望太快地高估或低估这些预测值。但若对于一个时间序列的随机变异极小，则选择较大的加权系数，以便当预测误差发生时，能根据其变化的状况，很快降低预测值或升高预测值。选择加权系数的标准亦如前面移动平均计算选择期数的标准一样，即要选择一个 α 值使均方误差（MSE）为最小。

以指数平滑法计算前例的预测值，取 $\alpha = 0.2$，计算结果如表 15-4 所示。

表 15-4　指数平滑法的预测值

周　　次	时间序列值 Y_t	移动平均预测值 $\hat{Y}_t$	预测误差 $Y_t-\hat{Y}_t$	预测误差平方$(Y_t-\hat{Y}_t)^2$
1	63			
2	81	63	18	324
3	72	66.6	5.4	29.16
4	63	67.68	–4.68	21.9
5	54	66.74	–12.74	163.31
6	72	64.19	7.81	61
7	87	65.75	21.25	451.56
8	84	70	14	196
9	60	72.8	–12.8	163.84
10	48	70.24	–22.24	494.62
11	60	65.79	–5.79	33.52
12	66	64.63	1.37	1.88
13		64.9		$\sum=1\,940.79$

15.5　用趋势方程预测时间序列

在之前描述自变量 X 与因变量 Y 之间的线性关系的估计回归方程为：

$$\hat{Y}=b_0+b_1X$$

而在预测时，为了使自变量为时间的事实更明显，以 t 代替上式中的 X，另外以 T_t代替 $\hat{Y}$。因此估计销售量的线性趋势可表示成如下的时间函数：

线性趋势的方程：$T_t=b_0+b_1t$。

其中：

T_t为第 t 期的时间序列预测值（以趋势为准）。

b_0 为趋势线的截距。

b_1 为趋势线的斜率。

t 为时点。

在上式中，令 t=1 表示时间序列数据第一个实际值所对应的时间，t=2 为第二个观察值所对应的时间等。至于估计回归系数（b_0 与 b_1）的计算公式在以前已经提过了，再重述如下，并以 t 代替 X，Y_t代替 Y_i。

斜率（b_0）与截距（b_0）的计算公式如下：

$$b_1=\frac{\sum tY_t-\left(\sum t\sum Y_t\right)/n}{\sum t^2I-\left(\sum t\right)^2/n}$$

$$b_0=\overline{Y}-b_1\overline{t}$$



其中：

Y_1 为第 t 期的时间序列实际值。

n 为期数。

$\overline{y}$ 为时间序列的平均值，$\overline{y}=\sum Y_t/n$。

$\overline{t}$ 为时点的平均值，$\overline{t}=\sum t/n$。

15.6　预测含趋势与季节成分的时间序列

在 15.5 节说明了如何预测含趋势成分的时间序列。本节将讨论如何预测含趋势与季节两种成分的时间序列，所使用的方法是由时间序列中先除去季节效应，此步骤称为剔除季节性。在剔除季节性后，时间序列将仅含趋势成分，然后就可用 15.5 节介绍的方法，辨认其趋势成分。而后应用趋势估计模型，将可预测未来时期的时间序列的趋势成分。最后再以季节指数调整趋势估计值。如此一来，将可辨认趋势与季节成分，并在预测时同时考虑这两者。

除了趋势成分（T）与季节成分（S）之外，将假设该时间序列也有不规则成分（I）。不规则成分是说明无法由趋势与季节成分解释的任何随机效应。以 T_t、S_t及 I_t指明时间 t 的趋势、季节与不规则成分，将假设时间序列模型（multiplicative time series model）表示为：

$$Y_t=T_t\times S_t\times I_t$$

在此模型中，T_t是用绝对数代表趋势，S_t与 I_t则是以相对数值度量，若其值高于 1.00，则表示效应在趋势之上；若其值低于 1.00，则表示效应在趋势的下方。

1．剔除时间序列的季节性

求季节指数的目的通常是要剔除时间序列中的季节效应，此过程称为剔除时间序列的季节性。像当前商业调查与华尔街日报等刊物常报导经季节变异调整过后的经济时间序列（除去季节性的时间序列）。利用乘法模型的符号，可得到：

$$Y_t=T_t\times S_t\times I_t$$

将各时间序列观察值除以对应的季节指数，即可将季节效应除去。

2．剔除季节性的时间序列估计趋势

当已有剔除季节性后的数据，可以直接利用这些数值来计算趋势。因此估计量的线性趋势方程，可以写成如下的时间函数：

$$T_t=b_0+b_1t$$

参数估计的具体方法参见 15.5 节。

3．循环成分

在数学上，可将乘法模型推广为如下含循环成分的模型：

$$Y_t = T_t \times C_t \times S_t \times I_t$$

循环成分与季节成分相同，也用趋势的百分比来表示，此成分可归结为时间序列的多年循环。与季节成分相类似，但经过的时间较长，很难收集足够的相关数据以估计循环成分。

15.7 利用回归模型预测时间序列

在回归分析讨论中，说明如何用一个或一个以上的自变量预测单一因变量的值。当回归分析被视为预测工具，可将要预测的时间序列值视为因变量。因此，若能找到一组良好的自变量，可建立预测时间序列的估计回归方程。

在建立估计回归方程时，需要一个包含因变量及所有自变量的观察值样本，而在时间序列分析中，N 个时期的时间序列数据，恰可作为用于此分析中的每一个变量的 N 个观察值的样本。对含有 k 个自变量的函数而言，以下列符号表示：

Y_t 为第 t 期时间序列的实际值。

X_{1t} 为第 t 期的第 1 个自变量值。

X_{2t} 为第 t 期的第 2 个自变量值。

⋮

X_{kt} 为第 t 期的第 k 个自变量值。

可以想象到，在一个预测模型中，自变量的选择有许多种，其中一种可能的选择是以时间为自变量。例如以时间为自变量，构建线性函数来估计该时间序列趋势。令 $X_{1t}=t$ 则可求得形式为：

$$\hat{Y}_t = b_0 + b_1 t$$

的估计回归方程，其中 $\hat{Y}_t$ 为时间序列 Y_t 值的估计值，而 b_0 与 b_1 为估计回归系数。在更复杂的模型中，可加入时间的高次幂项。例如，令：

$$X_{2t} = t^2$$

且

$$X_{3t} = t^3$$

则回归方程变成：

$$\hat{Y}_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + b_3 X_{3t}$$
$$= b_0 + b_1 t + b_2 t^2 + b_3 t^3$$

注意，此模型可提供具有曲线时间特征的时间序列预测值。

回归方法能否提供一个良好的预测值，依赖于所得到的自变量数据是否与此时间序列有紧密的关系。一般在建立一个估计回归方程时，会考虑到许多种自变量的组合。所以回归分析的部分程序，即将注意力集中于所要选择的自变量上，以期能提供一个最好的预测

模型。

因果预测模型利用与所预测的序列相关的时间序列，说明了时间序列行为的因果。回归分析是常用的建立这些因果模型的工具；相关的时间序列被看作自变量，而要预测的时间序列则是因变量。

另一种以回归为基础的预测模型，则其自变量为这一时间序列的所有前期值。例如，若以 $Y_1, Y_2,\cdots,Y_n$ 表示时间序列值，而因变量为 Y_t 则可能试图建立 $\hat{Y}_t=b_0+b_1Y_{t-1}+b_2Y_{t-2}+b_3Y_{t-3}$，对 Y_{t-1}，Y_{t-2} 等近期时间序列值估计回归方程。若以最近三期为自变量，则其估计回归方程为：

$$\hat{Y}_t=b_0+b_1Y_{t-1}+b_2Y_{t-2}+b_3Y_{t-3}$$

以时间序列的前期值为自变量的回归模型称为自回归模型（autoregressive model）。

最后，另一种以回归为基础的预测方法则综合前述所讨论的自变量。例如，可能选择时间变量、一些经济及人口统计变量与一些前期值加入自变量中。

15.8　其他预测模型

所谓简算法（naive method），是指一种不需依靠繁琐的计算和复杂的理论即可由历史数据得出预测值的方法，由于较其他预测方法简单、快速，故不失为一种有用的预测法。通常有下列两种方法，参见 Martin And Witt （1989a）。

方法一：$\hat{Y}_{t+1}=Y_t$

方法二：$\hat{Y}_{t+1}=Y_t\times\left(1+\dfrac{Y_t-Y_{t-1}}{Y_{t-1}}\right)$

其中 Y_t 为一组时间序列数据；$\hat{Y}_{t+1}$ 代表未来一期的预测值。

在方法一中，第 t+1 期的预测值即等于第 t 期的观测值。可以了解，在没有特殊状况的情形下，用简算法做预测，是直接而合理的，但是在使用这种方法时，通常要辅以其他方法。

在方法二中，第 t+1 期的预测值即等于第 t 期的观测值加上第 t 期的观测值乘以第 t 期的成长率。这样的方法，考虑了第 t 期的成长趋势对第 t+1 期的影响，一般来说，对稳定成长的时间序列做预测，用这样的方法并无不妥。若第 t 期有正成长，第 t+1 期便同样是正成长；若第 t 期正成长，第 t+1 期便为同幅度的正成长。

15.9　单变量时间序列预测模型

假设随机变量 Y_t 为在时间 t 的一个观测值，那么一组 Y_t 所成的序列，就称为一个时间序列。有所谓的 ARIMA 模型（autoregressive integrated moving average model），记作

Y_t～ARIMA（p, d, q），其公式如下：

$$\phi_p(B)Z_t = \theta_q(B)a_t$$

其中：

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

$$Z_t = (1-B)^d Y_t$$

滞后算子（backward shift operator）B，其公式如下：

$$BZ_t = Z_{t-1}$$

$$B^m Z_t = Z_{t-m}$$

Z_t为 t 时的观察值。

滞后差分算子（backward difference）$1-B$，其公式如下：

$$\overline{V}Z_t = Z_t - Z_{t-1} = (1-B)Z_t$$

相加运算（summation operation）S，其公式如下：

$$S = \overline{V}^{t-1}$$

$$\sum_{j=0}^{\infty} Z_{t-j} = Z_t + Z_{t-1} + Z_{t-2} + \cdots$$

$$= (1 + B + B^2 + \cdots)Z_t$$

$$= (1-B)^{t-1}Z_t = SZ_t = \overline{V}^{t-1}Z_t$$

白噪声（white noise）：$a_t, a_{t-1}, \ldots, a_{t-k}, \ldots$，其公式如下：

$$E(a_t) = 0，\quad V(a_t) = \sigma_a^2$$

1．自回归模型（autoregressive model，AR Model）

自回归模型的推导公式如下：

$$\tilde{Z}_t = Z_t - u$$

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + \cdots + \phi_p \tilde{Z}_{t-p} + a_t$$

$$\tilde{Z}_t = \phi_1 B\tilde{Z}_t + \phi_2 B^2 \tilde{Z}_t + \cdots + \phi_p B^p \tilde{Z}_t + a_t$$

所以 $a_t = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\tilde{Z}_t = \phi(B)\tilde{Z}_t$

$$\phi(B)\tilde{Z}_t = a_t \Leftrightarrow \tilde{Z}_t = \phi(B)a_t$$

$$\phi(B) = \phi^{-1}(B)$$

2．移动平均模型（moving average process model，MA Model）

移动平均模型的推导公式如下：

$$\tilde{Z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

$$\tilde{Z}_t = a_t - \theta_1 B a_t - \theta_2 B^2 a_t - \cdots - \theta_q B^q a_t$$

$$= a_t(1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \cdots - \theta_q B^q)$$

所以　$$\tilde{Z}_t = \theta(B) a_t$$

3．AR-MA 模型（mixed autoregressive moving average model Mixed AR-MA Model）

AR-MA 模型的推导公式如下：

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + \cdots + \phi_p \tilde{Z}_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

$$\tilde{Z}_t - \phi_1 B \tilde{Z}_t - \phi_2 B^2 \tilde{Z}_{t-2} - \cdots - \phi_p B^p \tilde{Z}_t = a_t - \theta_1 B a_t - \theta_2 B^2 a_t - \cdots - \theta_q B^q a_t$$

$$\phi(B)\tilde{Z}_t = \theta(B) a_t$$

一般 AR-MA 模型的 p、q 值都小于 2，$p,q \leqslant 2$。

4．季节循环性时间序列模型（seasonal autoregressive integrated moving average model，SARIMA Model）

有些时间序列有季节循环的特性，称为 SARIMA 模型，记作 Y_t～SARIMA$(P.D.Q)S$，其公式如下：

$$\phi_P(B) Z_t = \Theta_Q(B) a_t$$

其中：

$\phi_P(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_P B^{PS}$。

$\Theta_P(B) = 1 - \Theta_1 B - \Theta_2 B^2 - \cdots - \Theta_Q B^{QS}$。

S 为季节循环期数，$Z_t = (1 - B^S)^D Y_t$。

时间序列模型是依照随机变量间的相关性而建立的，若是有外在的因素干扰，则时间序列趋势必有所改变，鉴于此，在做时间序列分析时，可以考虑干扰因子模型，其公式如下：

$$Y_t = \frac{\omega(B) B^b}{\delta(B)} I_t + N_t$$

其中：

N_t 为单变量时间序列模型。

$$I_t = S_t = \begin{cases} 0 & \text{干扰因子发生前} \\ 1 & \text{干扰因子发生后} \end{cases}$$

在应用时间序列分析方法时，最重要的假设是这个序列的平稳性。但是在实际应用方面，许多时间序列都不符合平稳的要求，针对这个问题，有两个解决方法：一是对 Y_t 做方差平稳转换（variance stabilizing transformation）；二是对 Y_t 做差分。在实际应用时，应该先决定是否要做方差平稳转换，其次再决定如何做差分。模型中的 a_t 表示残差项，如果模型配置良好，残差项应该像是一个白噪声过程。单变量时间序列模型的建立过程主要有三个阶段：模型识别（identification）、参数估计（estimation）和模型诊断（diagnostic checking），

如图 15-8 所示。当模型诊断时发现拟合不佳，应注意拟合不佳的模型有何特征，以便决定其他可能的模型，此时再重复建立模型的三个阶段。这样的过程是不断重复的，直到找出拟合好的模型为止。

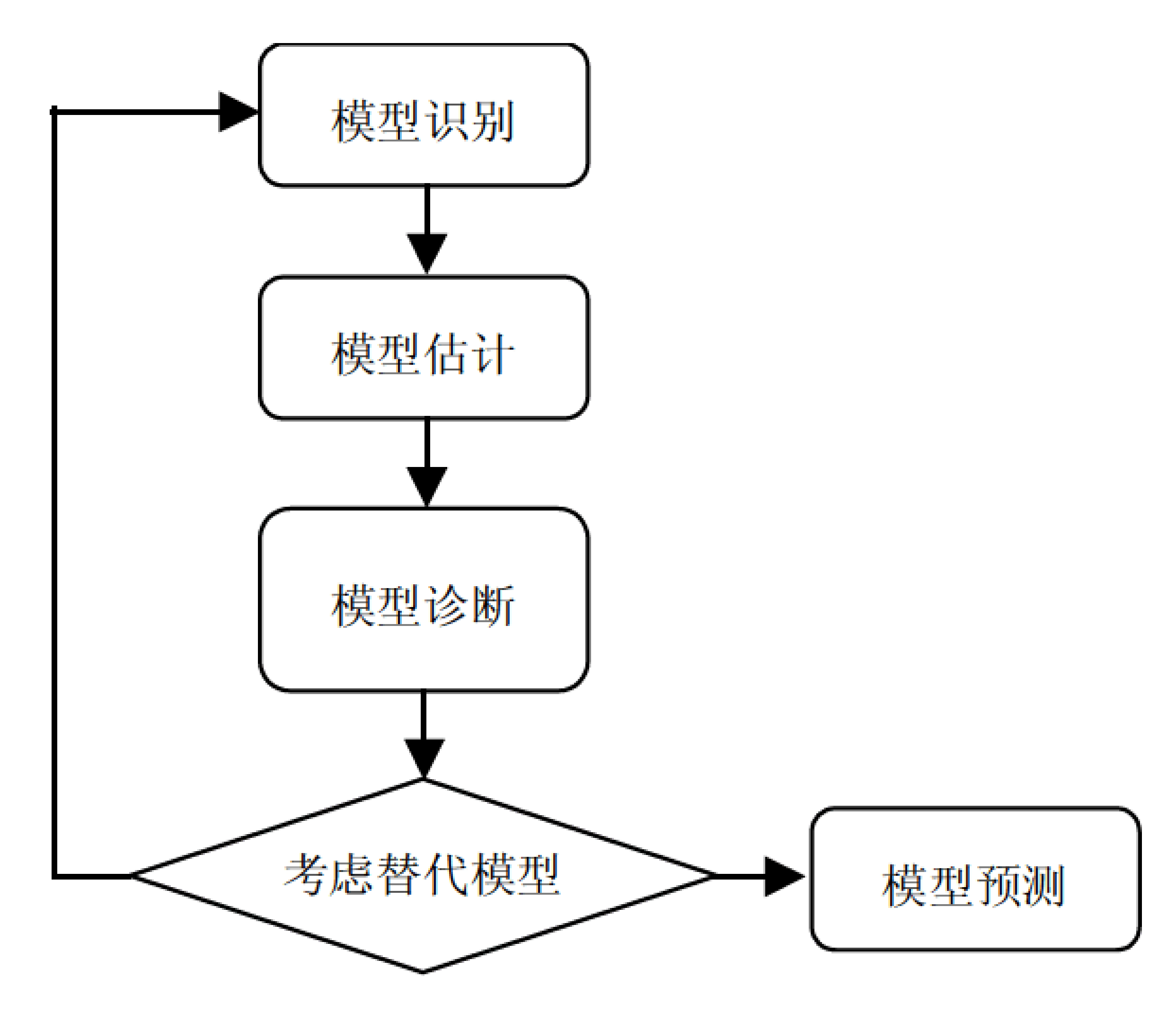


图 15-8　模型建立过程

15.10　时间趋势预测模型

时间趋势模型（trend curve analysis）是以需求量为被解释变量而以时间为解释变量，依据各种组合模型试图拟合最佳模型来表示需求量与时间之间的关系。这 10 种函数关系如下所示（Martin And Witt，1989）。

① 线性函数（linear）：$\hat{Y}_t = \beta_0 + \beta_1 t + \varepsilon_t$

② 双曲线函数（hyperbolic）：$\hat{Y}_t = \beta_0 + \beta_1 t^{t-1} + \varepsilon_t$

③ 限制型双曲线函数（constrained hyperbolic）：$1/\hat{Y}_t = \beta_0 + \beta_1 t^{t-1} + \varepsilon_t$

④ 变形双曲线函数（modified hyperbolic）：$1/\hat{Y}_t = \beta_0 + \beta_1 t + \varepsilon_t$

⑤ 指数函数（exponential）：$\ln \hat{Y}_t = \beta_0 + \beta_1 t + \varepsilon_t$

⑥ 变形指数函数（modified exponential）：$\ln \hat{Y}_t = \beta_0 + \beta_1 t^{t-1} + \varepsilon_t$

⑦ 半对数函数（semilog）：$\hat{Y}_t = \beta_0 + \beta_1 \ln t + \varepsilon_t$

⑧ 几何函数（geometric）：$\ln \hat{Y}_t = \beta_0 + \beta_1 \ln t + \varepsilon_t$

⑨ 二次函数（quadratic）：$\hat{Y}_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$

⑩ 对数二次函数（log quadratic）：$\ln \hat{Y}_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$

其中 $\hat{Y}_t$ 为 t 期的需求量，β_0，β_1，β_2 为参数，ε_t 为随机干扰项。

上述模型是一般形式的回归模型，因此在参数估计方面是以最小二乘法估计。在模型选取时，以模型解释度的高低及参数估计值显著与否进行筛选，再综合考虑模型的预测能力。实际应用时，用 Adj R-Square 值反映模型解释能力和 MAPE、RMSPE 评估其预测能力，选取一个最适合的理想模型。

15.11 Excel 2007 时间序列

SQL 2005 中的时间序列与一般所熟知的时序方法不尽相同，它是使用线性回归决策树的方法来分析时间相关的数据。它建立的模型可用来预测未来时刻的因变量值。Excel 2007 时间序列的操作步骤如下。

Step1：数据是三个地区 2001—2004 年 M200 型的销售记录，数据选取后单击【高级】下的【创建挖掘模型】按钮，如图 15-9 所示。

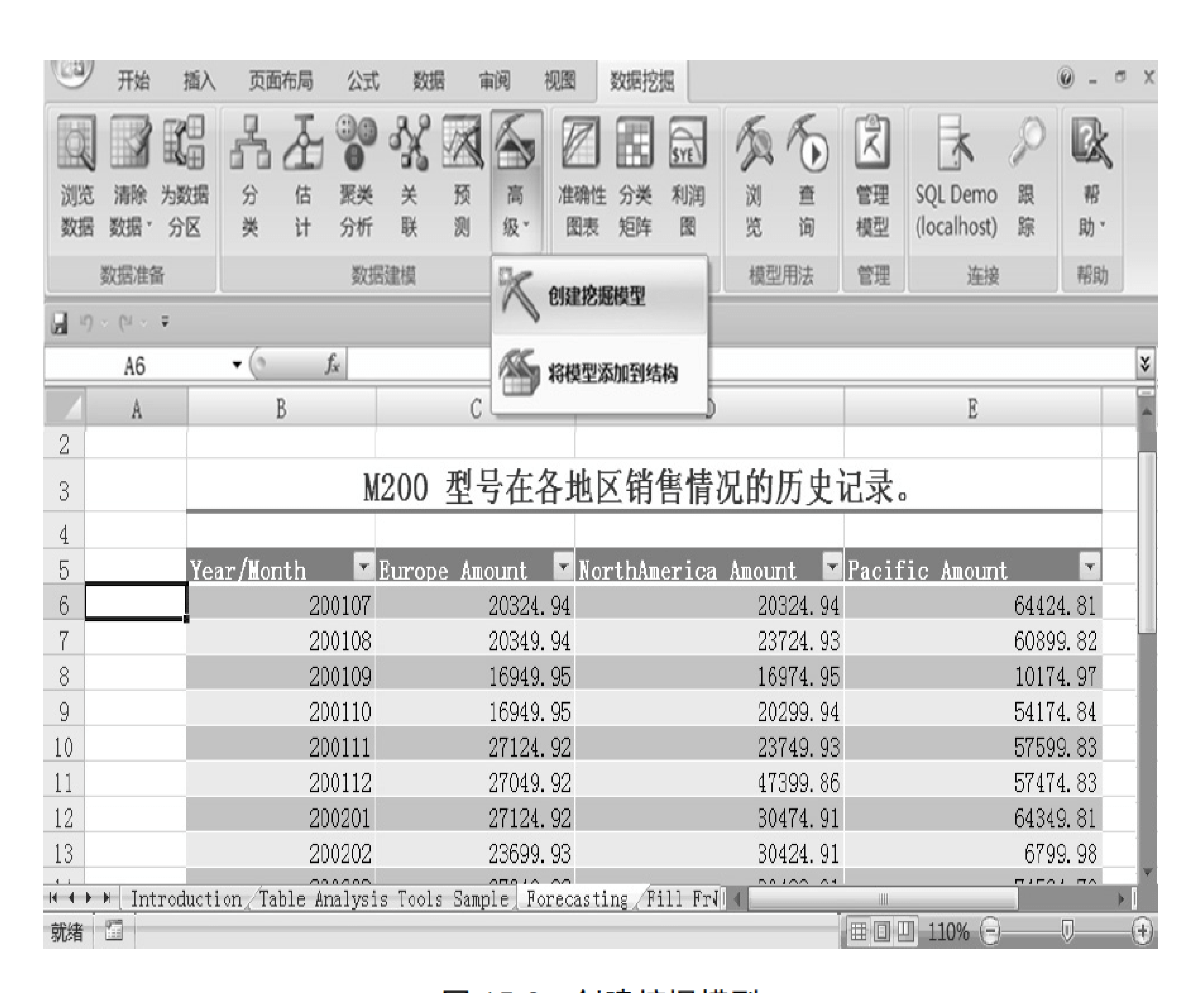


图 15-9 创建挖掘模型

Step2：在如图 15-10 所示的【创建模型向导入门】窗口中，单击【下一步】按钮。



图 15-10 【创建模型向导入门】窗口

Step3：选择挖掘算法，在【算法】下拉列表框中选择【Microsoft 时序】，如图 15-11 所示，单击【下一步】按钮。



图 15-11 选择挖掘算法

Step4：在如图 15-12 所示的【选择列】窗口中选择变量，将三个地区的变量都选为预测变量，单击【下一步】按钮。



图 15-12　选择变量

Step5：显示时间序列的决策树，如图 15-13 所示发现一共分为两层，以每月收入总额 200 305.859 作为分类水平。

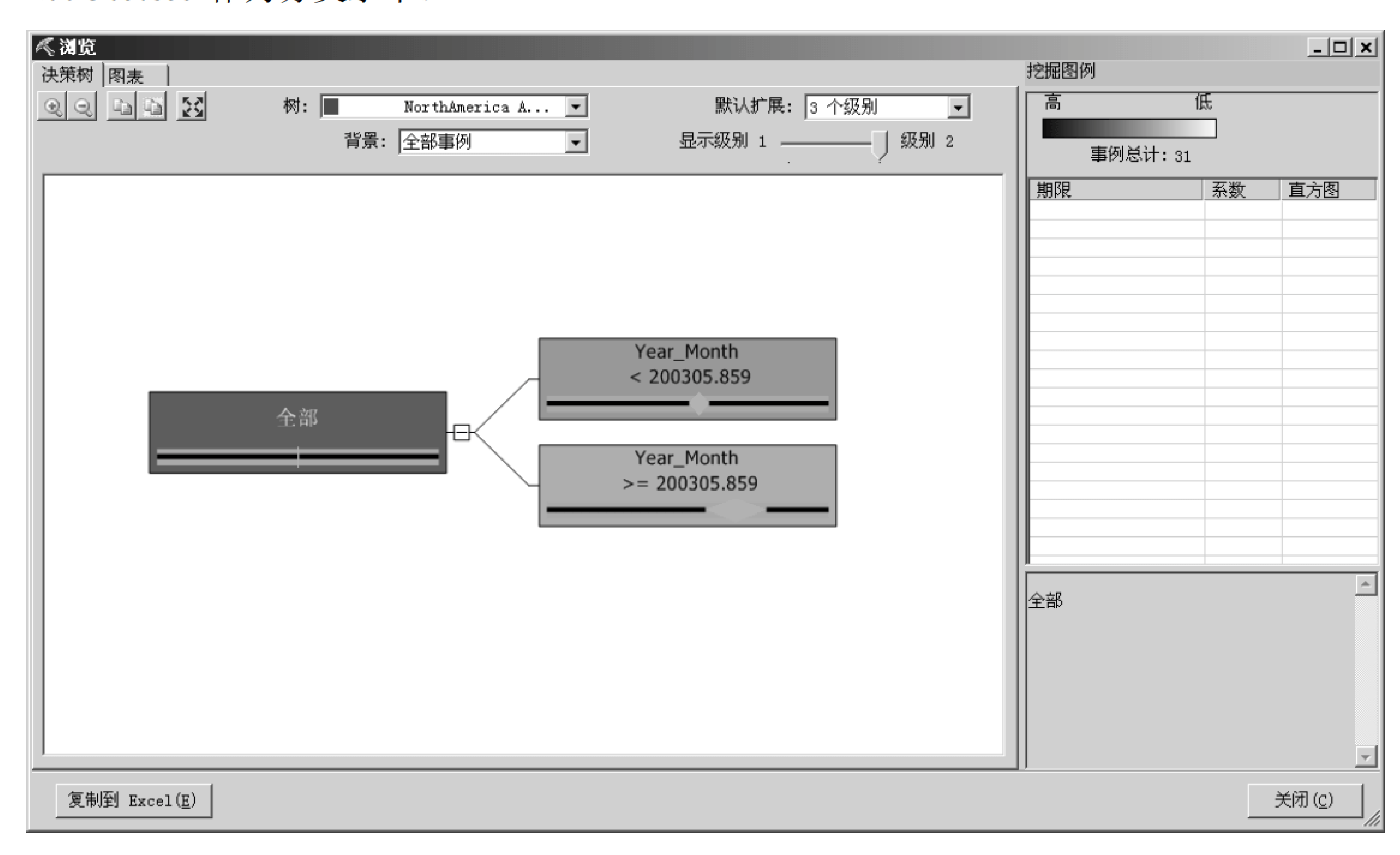


图 15-13　显示时间序列的决策树

Step6：将图表复制到 Excel，如图 15-14 所示。

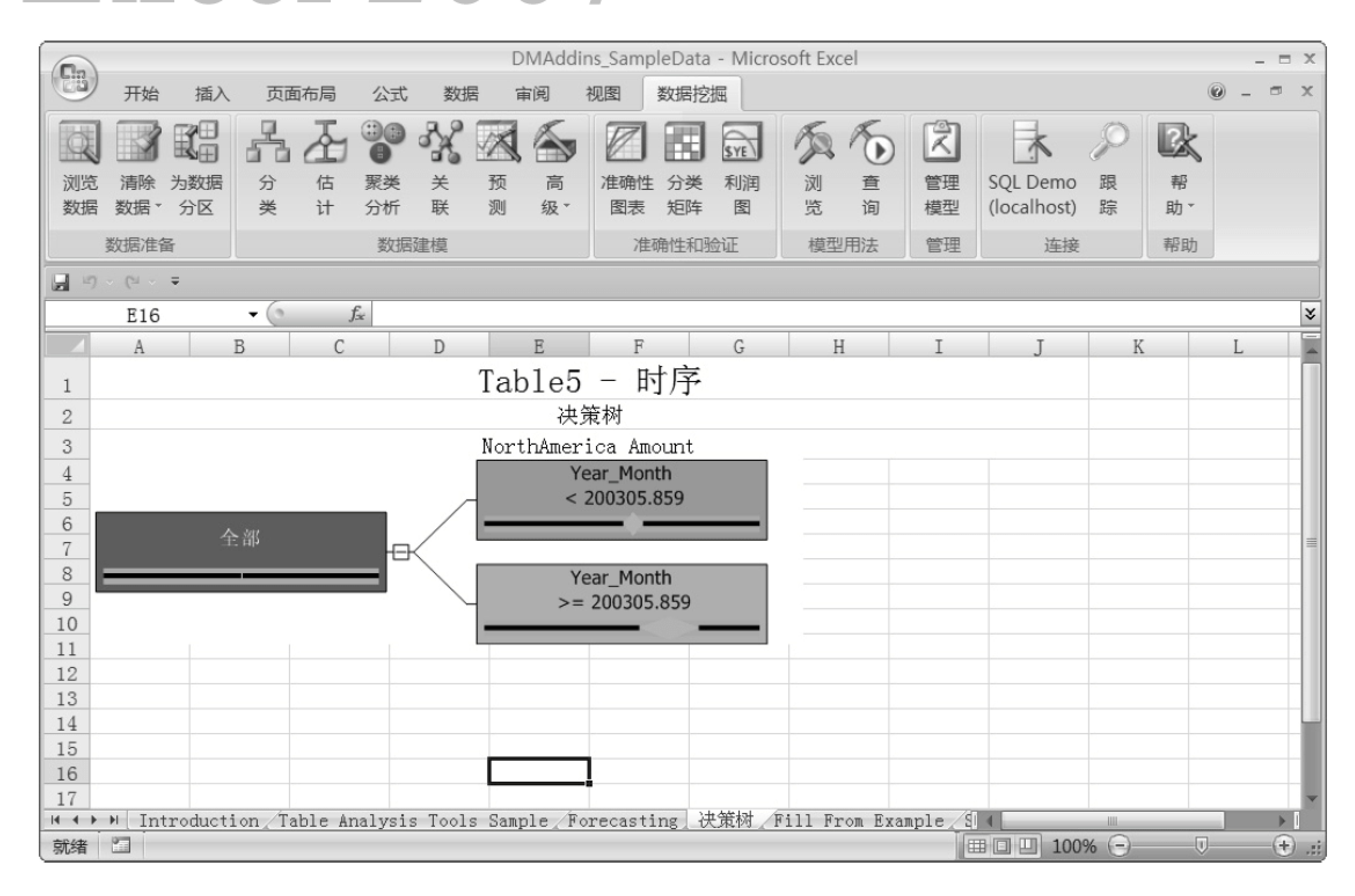


图 15-14　复制到 Excel

Step7：图 15-15 为三个地区的时间序列预测趋势图，在图中【预测步骤】微调按钮处可选择期望预测期数，在此选择预测 5 期，由图表可发现未来预测销售呈逐渐上升趋势。

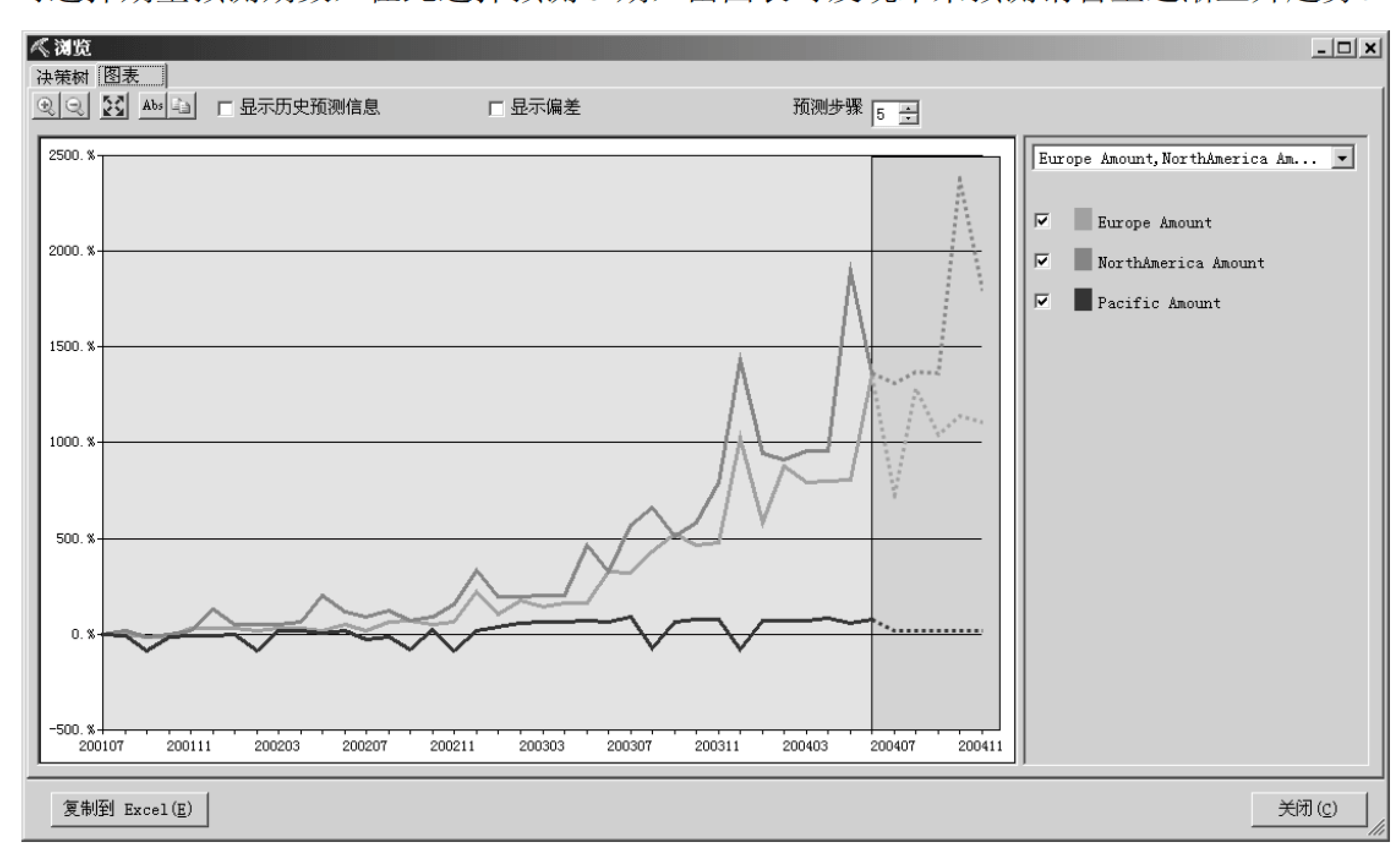


图 15-15　三地区时间序列预测趋势

Step8：将图表复制到 Excel，如图 15-16 所示。

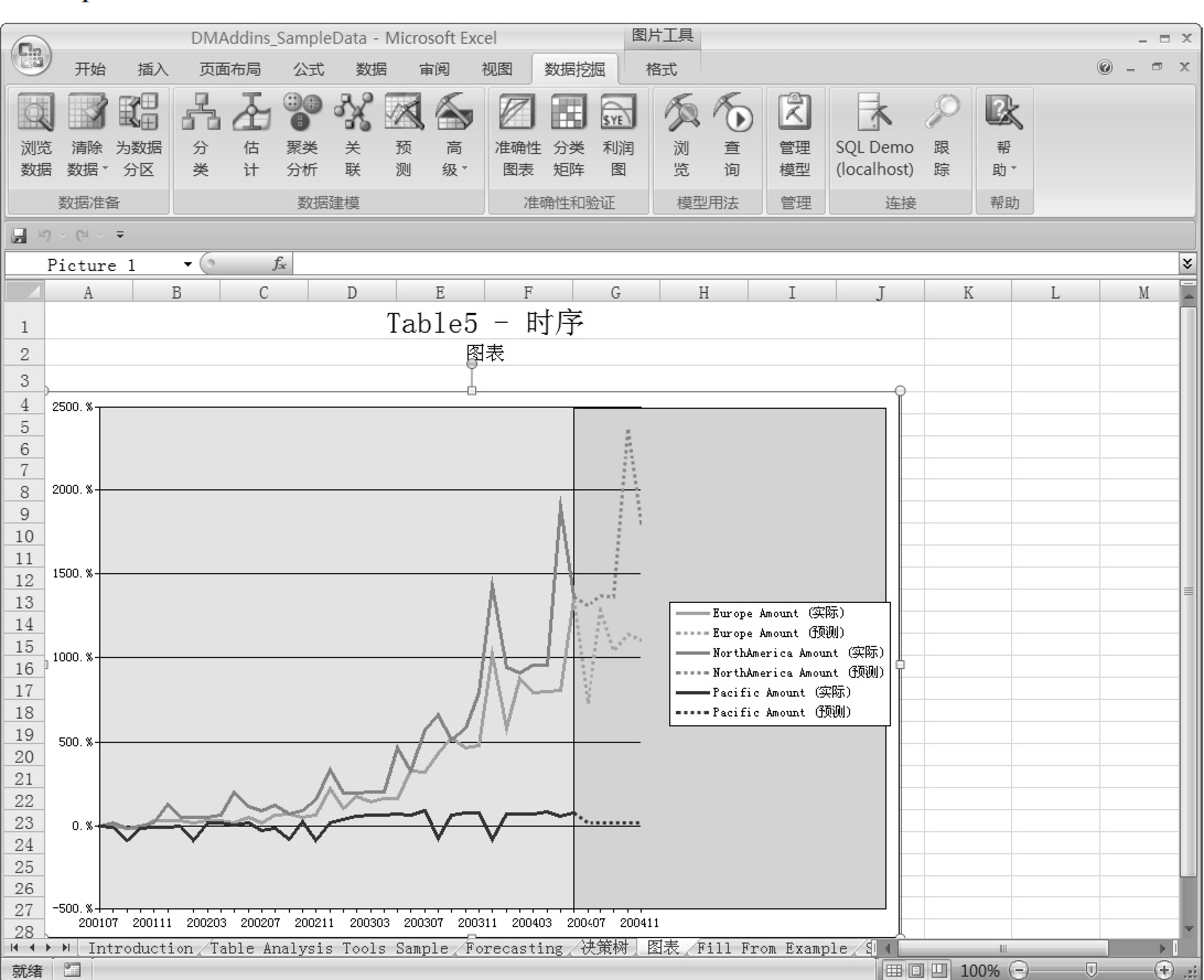


图 15-16　复制到 Excel

第 16 章　DMX 介绍

16.1　DMX 介绍

DMX，全名 data mining extension，在 SQL Server 2005 中用于建立和操作数据挖掘模型的语言，可以使用 DMX 建立新数据挖掘模型的结构，训练这些模型，以及浏览、管理与预测模型。DMX 是由数据定义语言（data definition language，DDL）语句、数据操作语言（data manipulation language，DML）语句，以及函数和操作符等所组成。使用前须定义一些对象如下。

❑ 标识符

定义名称对象，例如挖掘模型、挖掘结构和数据行。基本上分为两种，一般标识符与分隔标识符。一般标识符长度不可超过 100 字符，起始字符必须为下划线或被 Unicode Standard 2.0 所定义的字母。标识符不能为保留关键词，不论大小写，且中间不能有空格；分隔标识符以[ ]括住，在条件未符合一般标识符时使用；长度不可超过 100。

❑ 数据类型

定义挖掘模型数据行包含的数据类型。基本上有 text、long、boolean、double、date 五种数据类型。每种数据类型又分别支持不同内容类型，如定性变量、数值变量等。

❑ 表达式

通常包含单一或纯量值、对象，或数据表值的语法单位。

常数是代表单一特定值的符号。常数可以是字符串，或是数值或日期值。必须使用单引号“‘”来分隔字符串与日期常数。纯量函数会传回单一值，非纯量函数会传回数据表。而对象标识符在 DMX 中视为简单表达式。

❑ 操作符

配合一个或多个简单 DMX 表达式使用，以产生更复杂的 DMX 表达式。

❑ 函数

采用零、一或多个输入值，并传回纯量值或数据表的表达式。SQL 2005 中还可使用 VBA 或 Excel 的函数，也可以使用 common language runtime 程序设计语言建立扩充 DMX 功能的预存程序。

❑ 批注

文字元素，可以插入 DMX 语句或脚本中以说明语句的目的，方便程序撰写员未来开发或维护。//（双斜线）与--（双连字符）之后的所有文字将被视为批注，而符号/*、*/（斜线与星字号的配对），则是两者之中的文字将被视为批注。

❑ 保留关键词

保留给 DMX 使用的字，为数据库中的对象命名时不应使用这些字。若名称冲突时，

需使用标识符标记。

❑ 内容类型

定义挖掘结构数据行所包含的内容。每种算法支持不同的内容类型，基本上分为下列几种。

DISCRETE —— 如性别数据为一典型的离散数据。数据内包含有限的类别，即使是数值数据也不一定有排序意义，如电话号码。所有的数据类型皆可使用此种内容类型。

CONTINUOUS —— 数据为连续的数值数据，具有度量意义，可能有无限的小数值，如收入、身高等。date、double 和 long 三种数据类型支持此内容类型。

DISCRETIZED —— 数据为连续，但却须区分成分隔时，SQL 2005 会自动分隔成等份的值域，如身高从 150～180 可能有无限小数，分割成 150～160、161～170、171～180 三个值域。分割方式有 Automatic、CLUSTERS、EQUAL_AREAS、Thresholds 四种。date、double、long 和 text 四种数据类型支持此内容类型。

KEY——此数据行会唯一识别数据列。date、double、long 和 text 均支持此内容类型。

KEY SEQUENCE——为特定索引键类型，其值具有时间意义，而且已排序，不必为等距。double、long、text 和 date 支持此内容类型。

KEY TIME——为特定时间段索引键类型，其值代表已排序且会在某时段发生的值。double、long 和 date 支持此内容类型。

ORDERED —— 代表该数据为排序的值，如名次，但间距并没有意义，如第一名不代表成绩为第五名的五倍。所有数据类型，都支持此内容类型。

CYCLICAL——代表该数据具有循环且排序的值，如月份为一个典型例子。所有数据类型都支持此内容类型。

❑ 分布类型

定义数据的分布类型。定义之后，算法有可能得到更精确的结果。基本有三种模型，ormal 为正态分布、log normal 为对数正态分布、uniform 为均匀分布。

❑ 使用方式

在挖掘模型中须定义如何使用数据，基本类别如下：

Key 为索引键、Key Sequence 为具顺序性质的索引键、Key Time 为具时间性质的索引键、Predict 为同时用作输入与输出的值、PredictOnly 为只用作输出的值，其余未指定的值将是做输入值。

❑ 模型标识

定义其他的提示，如 Not null 为数据不能为空、REGRESSOR——可以在回归算法的回归公式里使用指定的数据行等。

16.2 DMX 函数介绍

基于挖掘阶段，大概分为三个阶段，下面分阶段介绍 DMX 应用。

16.2.1 模型建立

此阶段的代码编写存在一些语法习惯，例如：粗体为必须完全相同；斜体为使用者自定义的参数；|（竖线）在方括号或大括号内用来分隔语法项目，只能选择其中一种；[]（方括号）为选择性语法，使用时不输入方括号；{}（大括号）表示必要项目，使用时不输入大括号；,…指出逗号之前的项目可以重复多次，项目之间以逗号分割。

```
CREATE [SESSION] MINING MODEL <model>
    (
    [(<column definition list>)]
    )
USING <algorithm> [(<parameter list>)] [WITH DRILLTHROUGH]
```

model——该 model 的唯一名称。

SESSION——当连接关闭或会话超时时，建立的挖掘模型会自动移除。

algorithm——使用何种算法。

parameter list——定义算法的参数。

WITH DRILLTHROUGH——定义是否可以钻研。

column definition list——每行用逗号分隔，定义数据属性详细如下。

若为单一数据如下：

```
    <column name> <data type> [<Distribution>] [<Modeling Flags>] <Content
Type>
    [<prediction>] [<column relationship>]
```

若为巢状数据如下：

```
    <column name> TABLE [<prediction>] ( <non-table column definition
list> )
```

实际使用范例如下：

```
CREATE MINING MODEL PredictRisk
(ID KEY,
Gender TEXT DISCRETE,
Income LONG CONTINUOUS,
Job TEXT DISCRETE,
Area TEXT DISCRETE,
Risk TEXT DISCRETE PREDICT)
USING Microsoft_Decision_Trees
```

上述表明，使用微软决策树算法建立一个名称为 PredictRisk 的模型，包括六个数据行。其中 Risk 用作输入和输出的数据行，ID 字段为索引键，其余四个均为输入值。

16.2.2 模型训练

语法范例：

```
    INSERT  INTO  [MINING  MODEL]|[MINING  STRUCTURE]  <model>|<structure>
(<mapped model columns>) <source data query>

    INSERT INTO [MINING MODEL]|[MINING STRUCTURE]
    <model>|<structure>.COLUMN_VALUES (<mapped model columns>) <source data
query>
```

model——挖掘模型的名称。

structure——挖掘结构的名称。

mapped model columns——数据行标识符或巢状标识符的逗号分隔清单。

source data query——提供者自定义格式中的来源查询。

实例如下：

```
INSERT INTO PredictRisk
   (Id, Gender, Income, Job, Area, Risk)

   SELECT   ID, Gender, Income, Job, Area, Risk
FromCustomers
```

上两行表示插入的挖掘模型或挖掘结构，下两行表示查询的数据行以及对应的模型来源。

16.2.3 模型使用（预测）

基本语法如下：

```
SELECT [FLATTENED] [TOP <n>] <select expression list>
   FROM <model> | <sub select>
[NATURAL] PREDICTION JOIN <source data query>
   [ON <join mapping list>]
   [WHERE <condition expression>]
[ORDER BY <expression> [DESC|ASC]]
```

n——指定要返回数据列的个数，属性为整数。

select expression list ——从挖掘模型衍生数据标识符与表达式的分隔清单。

model——模型名称。

sub select——内嵌的 SELECT 语句。

source data query——来源查询。

join mapping list ——比较模型中的数据与来源查询中的数据的逻辑表达式。

condition expression——限制返回值的条件。

expression ——返回标量值的表达式。

实例如下：

```
SELECT NewCustomers.CustomerID, PredictRisk.Risk, CreditProbability
(PredictRisk)
FROMPredictRisk PREDICTION JOIN NewCustomers
ON  PredictRisk.Gender = NewCustomer.Gender
AND PredictRisk.Income = NewCustomer.Income
AND PredictRisk.Job = NewCustomer.Job
AND PredictRisk.Area = NewCustomer.Area
```

此外，若想删除挖掘模型或挖掘结构可使用：

```
DROP MINING MODEL <model >
DROP MINING STRUCTURE < structure>
```

若要将模型或结构输出或备份：

```
EXPORT <object type> <object name>[, <object name>] [<object type> <object
name>[, <object name] ] TO <filename> [WITH DEPENDENCIES]
```

实例如下：

```
EXPORT MINING MODEL [PredictRiskTO 'C:\PredictRisk.abf' WITH DEPENDENCIES
```

WITH DEPENDENCIES 指的是将所有相关的对象一起存入.abf 档案中，如数据来源和数据来源查看等。

同理，要将.abf 档案导入语法如下：

```
IMPORT [<object type> <object name>[, <object name>] [<object type>
<object name>[, <object name] ] ] FROM <filename>
```

实例如下：

```
IMPORT FROM 'C:\Predict.Risk.abf'
```

16.2.4 其他函数语法

BottomCount

根据次序表达式，以递增顺序返回一个数据表，并包含 count 数目的最底部数据行。

```
BottomCount(<table expression>, <rank expression>, <count>)
```

BottomPercent

类似 BottomCount，但是将 count 换成百分比，同样包含符合指定百分比表达式之最小数目的最底部数据行。

```
BottomPercent(<table expression>, <rank expression>, <percent>)
```

BottomSum

类似 BottomCount，但是将 count 换成 Sum，同样包含符合 sum 表达式之最小数目的

最底部数据行。

```
BottomSum(<table expression>, <rank expression>, <sum>)
```

TopCount

语法与功用类似 BottomCount，但是为递减顺序。

```
TopCount(<table expression>, <rank expression>, <count>)
```

TopPercent

语法与功用类似 BottomPercent，但是为递减顺序。

```
TopPercent(<table expression>, <rank expression>, <percent>)
```

TopSum

语法与功用类似 BottomSum，但是为递减顺序。

```
TopSum(<table expression>, <rank expression>, <sum>)
```

Cluster

返回最可能包含输入案例的聚类。不需参数，但只有当挖掘模型支持聚类时才可使用。

```
Cluster
```

ClusterProbability

类似 Cluster，返回输入案例属于聚类的概率。同样要挖掘模型支持聚类时才可使用。

```
ClusterProbability([<Node_Caption>])
```

IsDescendant

指出目前的节点是否从指定的节点衍生，返回一个布尔值。

```
IsDescendant(<NodeID>)
```

IsInNode

指出指定的节点是否包含案例，同样返回一个布尔值。

```
IsInNode(<NodeID>)
```

Lag

返回当前案例的日期与数据存在的最后日期之间的时间差。返回一个整数。

```
Lag()
```

Predict

在指定的数据行上执行预测。

```
Predict(<scalar column reference>, [option1], [option2], ,[INCLUDE_NODE_ID], n) Predict(<table column reference>, [option1], [option2], , [INCLUDE_NODE_ ID], n)
```

PredictAdjustedProbability

返回指定的可预测数据行的已调整概率。

```
PredictAdjustedProbability(<scalar column reference>, [<predicted state>])
```

PredictAssociation

预测各个数据列的关联性大小，可用于决策树、贝叶斯和类神经网络三种挖掘模型。

```
PredictAssociation(<table column reference>, option1, option2, n ...)
```

PredictCaseLikelihood

返回输入案例符合现有模型的概率。此函数只能配合聚类模型使用（聚类和时序聚类两种挖掘模型）。

```
PredictCaseLikelihood([NORMALIZED|NONNORMALIZED])
```

PredictHistogram

返回代表指定数据行的直方图的数据表。

```
PredictHistogram(<scalar column reference> | <cluster column reference>)
```

PredictNodeId

返回选取案例的 NodeID。

```
PredictNodeId(<scalar column reference>)
```

PredictProbability

返回指定数据行的概率。

```
PredictProbability(<scalar column reference>, [<predicted state>])
```

PredictSequence

预测顺序中的下一个值。

```
PredictSequence(<table column reference>)
PredictSequence(<table column reference, n>)
PredictSequence(<table column reference, n-start, n-end>)
```

PredictStdev

返回指定数据行的标准偏差。

```
PredictStdev(<scalar column reference>)
```

PredictSupport

返回指定数据行的支持值。

```
PredictSupport(<scalar column reference>, [<predicted state>])
```

PredictTimeSeries

返回时间序列的预测值。

```
PredictTimeSeries(<table column reference>)
PredictTimeSeries(<table column reference, n>)
PredictTimeSeries(<table column reference, n-start, n-end>)
```

```
PredictTimeSeries (<scalar column reference>)
PredictTimeSeries(<scalar column reference, n>)
PredictTimeSeries(<scalar column reference, n-start, n-end>)
```

PredictVariance
返回指定数据行的方差。

```
PredictVariance(<scalar column reference>)
```

RangeMax
探索指定的分割式数据行，返回预测分组的组距的最大数值。

```
RangeMax(<scalar column reference>)
```

RangeMid
探索指定的分割式数据行，返回预测分组的组距的中值。

```
RangeMid(<scalar column reference>)
```

RangeMin
探索指定的分割式数据行，返回预测分组的组距的最小数值。

```
RangeMin(<scalar column reference>)
```

16.3 DMX 数据挖掘语法

本节将针对 Microsoft SQL Server 2005 所提供的九种数据挖掘的方法论做参数介绍，并提供范例供读者参考。在分别介绍九种方法论的 DMX 数据挖掘语法前，先学习建立数据挖掘模型的基本语法。

```
CREATE [SESSION] MINING MODEL <model>
(
[(<column definition list>)]
)
USING <algorithm> [(<parameter list>)] [WITH DRILLTHROUGH]
CREATE MINING MODEL <model> FROM PMML <xml string>
```

其中各个自变量意义如表 16-1 所示。

表 16-1 自变量

自变量名称	描　　述
model	模型的唯一名称
column definition list	数据行定义的逗号分隔清单
algorithm	模型算法
parameter list	选择性，提供者自定义的算法参数的逗号分隔清单
XML string	XML 编码的模型（PMML），字符串必须使用单引号（'）括住，仅限高级使用

16.3.1 决策树

Microsoft 决策树算法可定义多个可能影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-2 所示。

表 16-2 决策树参数

参数名称	默认值	描述
MAXIMUM_INPUT_ATTRIBUTES	255	定义在使用功能选项之前，算法可以处理输入属性的数目；此值设定为 0 将关闭功能选项
MAXIMUM_OUTPUT_ATTRIBUTES	255	定义在使用功能选项之前，算法可以处理输出属性的数目；此值设定为 0 将关闭功能选项
SCORE_METHOD	3	决定用来计算分叉准则的方法。可用的选项：Entropy、Bayesian with K2 Prior 或 Bayesian Dirichlet Equivalent (BDE) Prior
SPLIT_METHOD	3	决定用来分叉节点的方法。可用的选项：Binary、Complete 或 Both
MINIMUM_SUPPORT	10	决定要在决策树中产生分叉所需的最小分叶案例数目
COMPLEXITY_PENALTY		控制决策树的成长。低值会增加分叉数目，而高值会减少分叉数目。默认值依据特定模型的属性数目而有所不同，如下列清单所述： ①1～9 个属性，默认值为 0.5 ②10～99 个属性，默认值为 0.9 ③100 个以上的属性，默认值为 0.99
FORCED_REGRESSOR		强制算法使用指定的数据行作为回归输入变量，不考虑算法计算出来的数据行的重要性。此参数只用于预测连续属性的决策树

[范例] 考虑性别、年龄、身份、收入、账户金额等属性，分类目标为信用评级（好、不好），决定顾客的信用评级。使用决策树分类建立的数据挖掘模型程序代码如下。

```
CREATE MINING MODEL Credit
(
[ID] LONG KEY,
[Sex] TEXT DISCRETE,
[Age] LONG DISCRETIZED,
[Identity] TEXT DISCRETE,
[Income] LONG DISCRETIZED,
[Accounting] LONG DISCRETIZED,
…
[CreditLevel] TEXT DISCRETE PREDICT
```

```
)
USING Microsoft_Decision_Trees(MAXIMUM_INPUT_ATTRIBUTES=0)
```

16.3.2 贝叶斯概率分类

Microsoft 贝叶斯概率分类算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-3 所示。

表 16-3 贝叶斯概率分类参数

参数名称	默认值	描述
MAXIMUM_INPUT_ATTRIBUTES	255	指定在使用功能选项之前，算法可以处理输入属性的最大数目；将此值设定为 0，会停用输入属性的功能选项
MAXIMUM_OUTPUT_ATTRIBUTES	255	指定在使用功能选项之前，算法可以处理输出属性的最大数目。将此值设定为 0，会停用输出属性的功能选项
MINIMUM_DEPENDENCY_PROBABILITY	0.5	指定介于输入和输出属性之间的最小相依概率。这个值会用来限制算法所产生内容的大小。此属性可设定为 0 到 1。越大的值会减少模型内容中的属性数目
MAXIMUM_STATES	100	指定算法所支持属性状态的最大数目。如果属性拥有的状态数目大于状态的最大数目，算法会使用属性最常用的状态并将其余的状态视为遗漏

[范例] 考虑性别、年龄、身份、收入四个属性，分类目标为办卡（会、不会），决定会员是否会办理信用卡。使用贝叶斯概率分类建立的数据挖掘模型程序代码如下。

```
CREATE MINING MODEL CreditCards
(
[ID] LONG KEY,
[Sex] TEXT DISCRETE,
[Age] LONG DISCRETIZED,
[Identity] TEXT DISCRETE,
[Income] LONG DISCRETIZED,
[UseCard] TEXT DISCRETE PREDICT
)
USING Microsoft_Naive_Bayes(MAXIMUM_INPUT_ATTRIBUTES=5)
```

16.3.3 关联规则

Microsoft 关联分析算法可以定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-4 所示。

表 16-4　关联分析参数

参数名称	默认值	描述
MINIMUM_SUPPORT	0.03	指定算法产生规则之前必须包含项目集的最小案例数目。将此值设定为小于 1，是以总案例数的百分比来指定最小案例数目；将此值设定为大于 1 的整数，是以必须包含项目集的绝对案例数目来指定最小案例数目。如果内存有限，算法可增加此参数的值
MAXIMUM_SUPPORT	1	指定项目集可支持的最大案例数目。如果此值小于 1，则此值代表总案例数的百分比；大于 1 的值代表可包含项目集的绝对案例数目
MINIMUM_ITEMSET_SIZE	1	指定项目集内所允许的最小项目数目
MAXIMUM_ITEMSET_SIZE	3	指定项目集内所允许的最大项目数目。将此值设定为 0，即代表项目集没有大小限制
MAXIMUM_ITEMSET_COUNT	200 000	指定要产生的最大项目集数目。如果没有指定数目，算法会产生所有可能的项目集
MINIMUM_PROBABILITY	0.4	指定规则为 True 的最小概率。例如，将此值设定为 0.5 是指定不产生概率小于 50%的规则
OPTIMIZED_PREDICTION_COUNT		定义要为预测进行快取或优化的项目数目

[范例]　考虑性别、年龄、收入、最喜欢的演员、最喜欢的导演、最喜欢的电影类型等属性，决定最有卖点的电影内容及其市场。使用关联规则建立的数据挖掘模型程序代码如下。

```
CREATE MINING MODEL GoodMovies
(
[ID] LONG KEY,
[Sex] TEXT DISCRETE,
[Age] LONG DISCRETIZED,
[Income] LONG DISCRETIZED,
[FavoriteActor] TEXT DISCRETE PREDICT,
[FavoriteDirector] TEXT DISCRETE PREDICT,
[FavoriteMovie] TEXT DISCRETE PREDICT
)
USING Microsoft_Association_Rules (MINIMUM_SUPPORT=0.05,
MINIMUM_PROBABILITY=0.70)
```

16.3.4　聚类分析

Microsoft 聚类算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-5 所示。

表 16-5 聚类算法参数

参数名称	默认值	描述
CLUSTERING_METHOD	1	指定算法要使用的聚类方法。可用的聚类方法有：可扩充的 EM、不可扩充的 EM、可扩充的 K-means 和不可扩充的 K-means
CLUSTER_COUNT	10	指定算法要建立的聚类数目。如果无法从数据建立聚类数目，则算法会尽可能建立最多的聚类。将 CLUSTER_COUNT 设定为 0 会造成算法使用启发法，对于建立的聚类数做出最好的决定
CLUSTER_SEED	0	指定在模型建立的初始阶段，用于随机产生聚类的种子
MINIMUM_SUPPORT	1	指定每一个聚类的最小观测数目
MODELLING_CARDINALITY	10	指定在聚类处理期间构建的范例模型数目
STOPPING_TOLERANCE	10	指定用来决定何时到达收敛状态以及算法完成建立模型的值。当聚类概率的整体变更小于本参数值除以模型大小的比率时，就到达收敛状态
SAMPLE_SIZE	50 000	指定如果 CLUSTERING_METHOD 参数设定为可扩充的聚类方法之一时，算法使用在每个进程上的观测数目。将本参数设定为 0 会导致将整个数据集在单一进程中聚类。这会造成内存和效能的问题
MAXIMUM_INPUT_ATTRIBUTES	255	指定使用功能选项之前，算法可以处理输入属性的最大数目。将此值设定为 0 即表示属性数目没有上限
MAXIMUM_STATES	100	指定算法所支持属性状态的最大数目。如果属性的状态数目大于状态数目上限，则算法会使用属性最常用的状态，而忽略其余状态

[范例] 以下范例以顾客的年龄与收入作为分群维度做聚类分析。

```
CREATE MINING MODEL Customer_Clustering
(
[ID] LONG KEY,
[Age] LONG DISCRETIZED,
[Income] LONG DISCRETIZED
)
USING Microsoft_Clustering (CLUSTERING_METHOD=3)
```

16.3.5 时序聚类

Microsoft 时序聚类算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-6 所示。

表 16-6　时序聚类算法参数

参 数 名 称	默 认 值	描　　述
CLUSTER_COUNT	10	指定算法要建立的聚类数目。如果无法从数据建立聚类数目，则算法会尽可能建立最多的聚类。将本参数值设定为 0，会导致算法使用启发式来判断可建立的最佳聚类数目
MINIMUM_SUPPORT	10	指定每一个聚类的最小案例数目
MAXIMUM_SEQUENCE_STATES	64	指定一个序列可以具有的最大状态数目。将此值设定为大于 100 的数字将可能导致算法建立一个无法提供有用信息的模型
MAXIMUM_STATES	100	针对算法支持的非序列属性指定最大状态数目。如果非序列属性的状态数目大于最大状态数目，算法会使用该属性最常用的状态，并将其余的状态视为遗漏

[范例]　考虑 Web 应用程序的用户经常以各种路径浏览网站，根据浏览站点的页面类型对用户进行分组，以帮助分析消费者并决定消费者可能的浏览网站，提高网站效益。

```
CREATE MINING MODEL WebSequence
(
[CustomerId] TEXT KEY,
[Location] TEXT DISCRETE,
[ClickPath] TABLE PREDICT
(
    [SequenceId] LONG KEY Sequence,
    [URLCategory] TEXT,
)

)
USING Microsoft_Sequence_Clustering (CLUSTER_COUNT=0)
```

16.3.6　线性回归

Microsoft 线性回归分析算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-7 所示。

表 16-7　线性回归分析参数

参 数 名 称	默 认 值	描　　述
MAXIMUM_INPUT_ATTRIBUTES	255	定义使用功能选项之前，算法可以处理输入属性的数目。如此值设定为 0 将关闭功能选项
MAXIMUM_OUTPUT_ATTRIBUTES	255	定义使用功能选项之前，算法可以处理输出属性的数目。如此值设定为 0 将关闭功能选项

续表

参数名称	默认值	描述
FORCED_REGRESSOR		强制算法使用指定的数据行作为回归输入变量，不考虑算法计算出来的数据行的重要性

[范例]　用身高预测体重，使用线性回归分析建立的数据挖掘模型程序代码如下。

```
CREATE MINING MODEL PreWeight
(
[Id] LONG KEY,
[Height] LONG DISCRETE,
[Weight] LONG DISCRETE PREDICT
)
USING Microsoft_Linear_Regression
```

16.3.7　Logistic 回归

Microsoft Logistic 回归算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-8 所示。

表 16-8　Logistic 回归参数

参数名称	默认值	描述
HOLDOUT_PERCENTAGE	30	指定用于计算测试错误的训练数据内的观测百分比，本参数在训练挖掘模型时是作为停止准则的一部分
HOLDOUT_SEED	0	在随机决定测试数据时，指定用来植入虚拟随机产生器的数字。如果本参数值设定为 0，则此算法会依据挖掘模型的名称产生种子，以保证在重新处理期间模型内容保持不变
MAXIMUM_INPUT_ATTRIBUTES	255	定义使用功能选项之前，算法可以处理输入属性的数目；如此值设定为 0 将关闭功能选项
MAXIMUM_OUTPUT_ATTRIBUTES	255	定义使用功能选项之前，算法可以处理输出属性的数目；如此值设定为 0 将关闭功能选项
MAXIMUM_STATES	100	指定算法所支持属性状态的最大数目。如果属性拥有的状态数目大于状态的最大数目，算法会使用属性最常用的状态，并忽略其余的状态
SAMPLE_SIZE	10 000	指定用来训练模型的观测数目。此算法提供者会使用此数字或不包括在测试百分比（由 HOLDOUT_PERCENTAGE 参数指定）中的总观测数的百分比，以较小者为准。换句话说，如果 HOLDOUT_PERCENTAGE 设定为 30，则算法将使用此参数的值，或等于总观测数 70% 的值，以较小者为准

[范例]　考虑有肥胖或抽烟情形的人会得高血压的人数。

```
CREATE MINING MODEL Logistic_Hypertension
(
[No] LONG KEY,
[Fat] Boolean DISCRETE,
[Smoke] Boolean DISCRETE,
[People] LONG DISCRETE,
[Hypertension] LONG DISCRETE PREDICT
)
USING Microsoft_Logistic_Regression
```

16.3.8　类神经网络

Microsoft 类神经网络算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-9 所示。

表 16-9　类神经网络参数

参数名称	默认值	描述
HIDDEN_NODE_RATIO	4.0	指定隐藏神经与输入和输出神经的比例。使用下列公式决定隐藏层中的初始神经数目 HIDDEN_NODE_RATIO * SQRT(Total input neurons * Total output neurons)
HOLDOUT_PERCENTAGE	30	指定用来计算测试错误训练数据内的观测百分比，这可作为训练挖掘模型时停止准则的一部分
HOLDOUT_SEED	0	在算法随机决定测试数据时，指定用来植入虚拟随机产生器的数字。如果此参数设定为 0，此算法会依据挖掘模型的名称产生种子，以保证在重新处理期间，模型内容保持不变
MAXIMUM_INPUT_ATTRIBUTES	255	决定在运用功能选项之前可提供给算法之输入属性的最大数目。将此值设定为 0，会停用输入属性的功能选项
MAXIMUM_OUTPUT_ATTRIBUTES	255	决定在运用功能选项之前可提供给算法之输出属性的最大数目。将此值设定为 0，会停用输出属性的功能选项
MAXIMUM_STATES	100	指定算法支持的每个属性之分隔状态的最大数目。如果特定属性的状态数目大于对这个参数所指定的数字，则算法会使用该属性最常用的状态，并将剩余状态视为遗漏

续表

参数名称	默认值	描述
SAMPLE_SIZE	10 000	指定用来训练模型的观测数目。此算法会使用此数字或不包括在测试数据中之总观测数的百分比（由 HOLDOUT_PERCENTAGE 参数指定），以较小者为准。换句话说，如果 HOLDOUT_PERCENTAGE 是设定为 30，则算法将使用这个参数的值或等于总观测数 70%的值，以较小者为准

[范例] 以性别、年龄、职业、教育程度、小孩数等属性作为输入变量，预测会员拥有的信用卡数。使用类神经网络建立的数据挖掘模型程序代码如下。

```
CREATE MINING MODEL CardNumber
(
[ID] LONG KEY,
[Sex] TEXT DISCRETE,
[Age] LONG DISCRETIZED,
[Occupation] TEXT DISCRETE,
[Education] TEXT DISCRETE,
[TotalChildren] LONG DISCRETIZED,
[OwnCard] LONG DISCRETE PREDICT
)
USING Microsoft_Neural_Network(HOLDOUT_PERCENTAGE=20)
```

16.3.9 时间序列

Microsoft 时间序列算法可定义多个会影响所产生的挖掘模型的效能和精确度的参数。具体的参数描述如表 16-10 所示。

表 16-10 时间序列参数

参数名称	默认值	描述
MINIMUM_SUPPORT	10	指定要在每一个时间序列树中产生分割所需时间配量的最小数目
COMPLEXITY_PENALTY	0.1	控制决策树的成长。减少此值可增加分割的可能性。增加此值则减少分割的可能性
PERIODICITY_HINT	{1}	提供算法关于数据周期性的提示。例如，若每年销售不同，序列中的度量单位是月，则周期性是 12。此参数采用{n[, n]}的格式，其中 n 是任何正数。方括号[]内的 n 是选择性的，可以视需要而重复

续表

参数名称	默认值	描述
MISSING_VALUE_SUBSTITUTION		指定缺失值替换的方法。依默认，数据中不允许有不规则的间距或不完全的边缘。以下是可用来填满不规则间距或边缘的方法：依据上一个（Previous）值、依据平均（Mean）值或依据特定数值常数（Numeric Constant）
AUTO_DETECT_PERIODICITY	0.6	指定 0 和 1 之间的数值，用来检测周期性。将这个值设定为越接近 1，就会探索更多接近周期性的模型，并自动产生周期性提示。处理大量周期性提示时，可能会造成更长的模型培训时间及更精确的模型。如果将此值设定为越接近 0，则只会检测到周期性很强的数据
HISTORIC_MODEL_COUNT	1	指定要建立的历程记录模型数目
HISTORICAL_MODEL_GAP	10	指定两个连续历程记录模型之间的时间延迟。例如，将此值设定为 g，会造成要建立历程记录模型的数据，按 g、2*g、3*g 等的间隔而产生时间配量截断

[范例] 假定政府要预测未来中国台湾地区人口总数，使用时间序列建立的数据挖掘模型程序代码如下。

```
CREATE MINING MODEL PopulationNumber
(
[Time] DATE KEY,
[Population] LONG DISCRETIZED PREDICT
)
USING Microsoft_Time_Series
```

16.4 DMX 应用范例

以下对 DMX 做较为完整的应用范例介绍，让读者能更清楚地知道 DMX 的用法。在本节的范例介绍中，以一般而言的数据挖掘所包含的五项功能：分类（classification）；估计（estimation）；预测（prediction）；关联分组（affinity grouping）；聚类（clustering），对各类分别举一范例做 DMX 语法介绍。需要注意的是，以下数据来源扩展名为“.xls”，请先使用 Microsoft SQL Server Management Studio 导入来源档案进入数据库。

16.4.1 分类

所谓分类（classification），即为按照分析对象的属性分门别类加以定义，建立类别（class）。例如，将信用卡申请者的申请结果区分为核卡或不核卡。使用的技巧有决策树（decision tree）等。以下举决策树技巧作为范例。

数据来源：投保.xls（导入成为数据库 Insure）

目标：

以保单号码（Policy No）为主键，缴费方式（Method）、保险型态 1（Insur_type4）、保险型态 2（Insurance_type）、性别（Rate_sex）、保额组别（Face_group）、理赔金组别（Claim_group）作为自变量，有无理赔（Cl_flag）作为预测变量进行分类，决定理赔行为判定。

模型建立：

```
CREATE MINING MODEL InsureDecisiontree
(
[Policy No] TEXT KEY,
[Insur_type4] TEXT DISCRETE,
[Insurance_type] TEXT DISCRETE,
[Rate_sex] TEXT DISCRETE,
[Face_group] TEXT DISCRETE,
[Claim_group] TEXT DISCRETE,
[Cl_flag] TEXT DISCRETE PREDICT
)
USING Microsoft_Decision_Trees
```

数据来源链接字符串：

```
    Provider=SQLNCLI.1; Data Source=DM-SERVER; Integrated Security=SSPI;
Initial Catalog=Insure
```

根据数据挖掘模型预测行为：

```
SELECT
  t.[face_group],[Insure].[Cl Flag]
From
  [Insure]
PREDICTION JOIN
  OPENQUERY([Insure],
    'SELECT
      [method],
      [insur_type4],
      [insurance_type],
      [rate_sex],
      [face_group],
```

```
        [claim_group],
        [cl_flag]
      FROM
        [dbo].[insure$]
      ') AS t
  ON
    [Insure].[Method] = t.[method] AND
    [Insure].[Insur Type4] = t.[insur_type4] AND
    [Insure].[Insurance Type] = t.[insurance_type] AND
    [Insure].[Rate Sex] = t.[rate_sex] AND
    [Insure].[Face Group] = t.[face_group] AND
    [Insure].[Claim Group] = t.[claim_group] AND
    [Insure].[Cl Flag] = t.[cl_flag]
```

16.4.2 估计

根据已有连续性数值相关属性数据，以获得某一属性的未知值。例如按照信用卡申请者的教育程度、行为类别来估计（estimation）其信用卡消费量。使用的技巧包括回归分析及类神经网络等。以下使用线性回归方法为例估计模型。

例如对一批投保数据建立回归模型，其中，保单号码（Policy No）为主键，保额（Face_amt）作为自变量，缴费年期（Collect_year）作为预测变量。

模型建立：

```
CREATE MINING MODEL InsureRegression
(
[Policy No] TEXT KEY,
[Face_amt] DOUBLE CONTINUOUS,

[Collect_year] DOUBLE CONTINUOUS PREDICT
)
USING Microsoft_Linear_Regression
```

数据来源链接字符串：

```
Provider=SQLNCLI.1; Data Source=DM-SERVER; Integrated Security=SSPI;
Initial Catalog=Insure
```

以下为本例的挖掘模型预测语法：

```
SELECT
  t.[collect_year]
From
  [Insure_R]
PREDICTION JOIN
  OPENQUERY([Insure],
    'SELECT
```

```
        [collect_year_ind],
        [collect_year],
        [face_amt]
      FROM
        [dbo].[insure$]
      ') AS t
ON
   [Insure_R].[Face Amt] = t.[face_amt] AND
  [Insure_R].[Collect Year] = t.[collect_year]
```

16.4.3 预测

根据对象属性的过去观察值来预测（prediction）该属性的未来值。例如由顾客过去的刷卡消费量预测其未来的刷卡消费量。使用的技巧包括回归分析、时间序列分析及类神经网络。以下使用时间序列分析为例预测模型。

以一批人口数据为例，预测明年人口水平。其中，年份（Year）为主键，15 岁以上人口总计（Population）作为输入及预测变量。

模型建立：

```
CREATE MINING MODEL Population_TimeSeries
(
[Year] LONG KEY TIME,
[Population] DOUBLE CONTINUOUS PREDICT
)
USING Microsoft_ Time_Series
```

数据来源链接字符串：

```
Provider=SQLNCLI.1; Data Source=DM-SERVER; Integrated Security=SSPI;
Initial Catalog=Population
```

如要进行人口预测，可使用以下语法做未来 5 年的人口预测：

```
SELECT PredictTimeSeries(Population,5) AS FuturePopulation
FROM Population_TimeSeries
```

16.4.4 关联分组

所谓关联分组（affinity grouping），就是在所有对象中，将相互关联的对象放在一起。例如超市中相关的盥洗用品（牙刷、牙膏、牙线）应放在同一货架上。在客户营销系统上，此种功能可用来确认交叉销售（cross selling）的机会以设计出吸引人的产品聚类从而增加销售。

对投保数据进行关联分组。其中，保单号码（Policy No）为主键，性别（Rate_sex）、缴费方式（Method）、保险形态 1（Insur_type4）、保险形态 2（Insurance_type）、通路

（Channel_code）、地区别（Company_code）作为输入变量，缴费方式（Method）、保险型态 1（Insur_type4）、保险型态 2（Insurance_type）、通路（Channel_code）、地区别（Company_code）作为预测变量。

模型建立：

```
CREATE MINING MODEL Insure_Association
(
[Policy No] TEXT KEY,
[Rate_sex] TEXT DISCRETE,
[Method] TEXT DISCRETE PREDICT,
[Insur_type4] TEXT DISCRETE PREDICT,
[Insurance_type] TEXT DISCRETE PREDICT,
[Channel_code] TEXT DISCRETE PREDICT,
[Company_code] TEXT DISCRETE PREDICT
)
USING Microsoft_Association_Rules (MINIMUM_PROBABILITY=0.60)
```

数据来源链接字符串：

```
Provider=SQLNCLI.1; Data Source=DM-SERVER; Integrated Security=SSPI;
Initial Catalog=Insure
```

在模型建立完成后，可以从内容中检索数据集及规则，以检索数据集 I 为例：

```
SELECT
Node_Description
FROM
Insure_Clustering.Content
WHERE
   Node_Type='I'
```

16.4.5 聚类

将异质总体中分割为较具同质性的聚类（clusters），相当于营销术语中的细分（segmentation），但是聚类分析事先并未对细分加以定义，而是利用某种算法细分数据。

对投保数据进行聚类分析。其中，保单号码（Policy No）为主键，将理赔件次（Claim_cnt），投保次数（Po_cnt）作为输入变量。

模型建立：

```
CREATE MINING MODEL Insure_Clustering
(
[Policy No] TEXT KEY,
[Claim_cnt] DOUBLE CONTINUOUS,
[Po_cnt] DOUBLE CONTINUOUS
)
USING Microsoft_Clustering
```

数据来源链接字符串：

```
   Provider=SQLNCLI.1; Data Source=DM-SERVER; Integrated Security=SSPI;
Initial Catalog=Insure
```

如需检索单一聚类内容，可使用以下语法来做查询，以检索聚类 1 为例：

```
SELECT t.* FROM Insure_Clustering
   NATURAL PREDICTION JOIN <Input Set> AS t
   WHERE Cluster()='聚类 1'
```



第 3 篇

其他分析工具介绍

- 分析关键影响因素
- 检测类别
- 从示例填充
- 预测
- 突出显示异常值
- 应用场景分析
- Visio 2007 数据透视分析

第 17 章　分析关键影响因素

分析关键影响因素工具可选取包含所要结果或目标值的数据行，然后分析数据集内的模型，以判断哪些因素对结果的影响力最强。例如，如果客户列表包含了会显示每一位客户在过去一年所购买的项目总计的数据行，则可以分析此数据表来判断哪些项目是客户购买最多的。

首先启动 Excel 2007 SQL 2005 DM addin 范例，在 Excel 数据表选项上右击，在弹出的快捷菜单中选择【自定义快速访问工具栏】命令，如图 17-1 所示。



图 17-1　Excel 2007 SQL 2005 DM addin 范例

在如图 17-2 所示的【Excel 选项】对话框中单击【加载项】按钮，并选择分析工具库，然后执行。

在如图 17-3 所示的【加载宏】对话框中，选中【分析工具库】复选框。

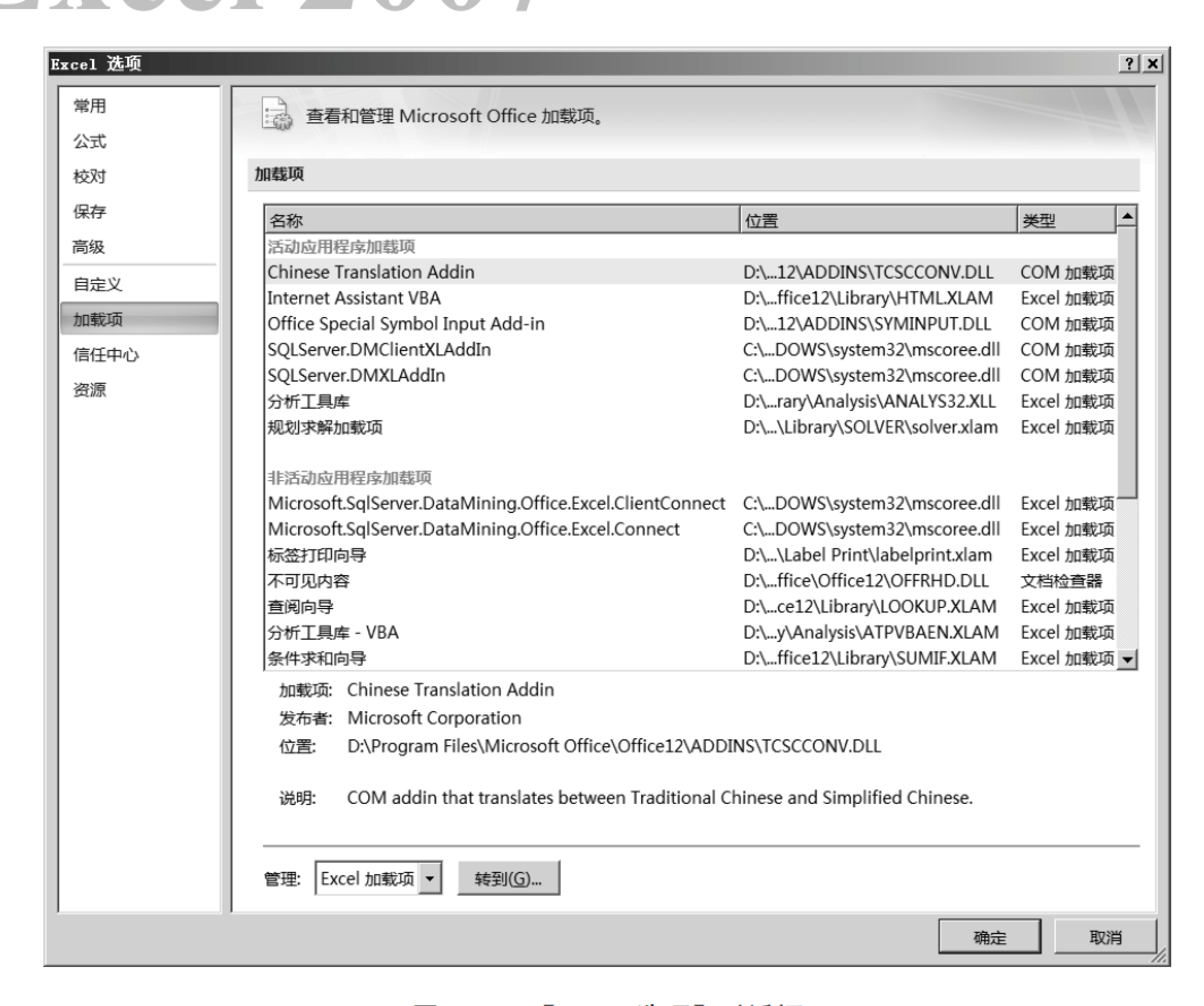


图 17-2 【Excel 选项】对话框

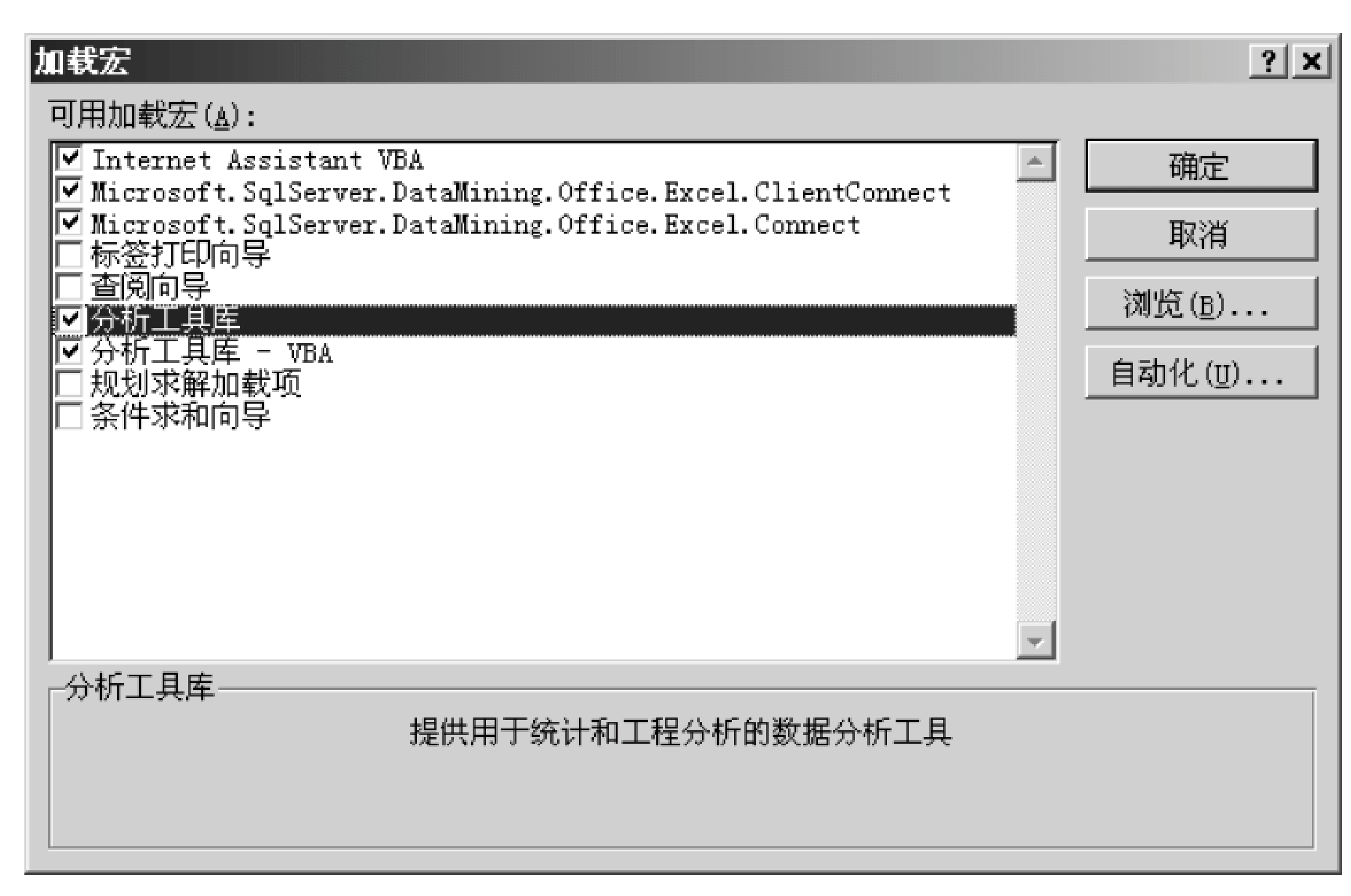


图 17-3 【加载宏】对话框

Microsoft Office Excel 会弹出一个对话框确认是否加载项，确定后 Microsoft Office 自动完成安装。

启动 Excel 2007 SQL 2005 DM addin 范例，选择所要分析的数据，如图 17-4 所示。

在 Excel 工具栏会出现【分析】和【设计】选项，如图 17-5 所示。

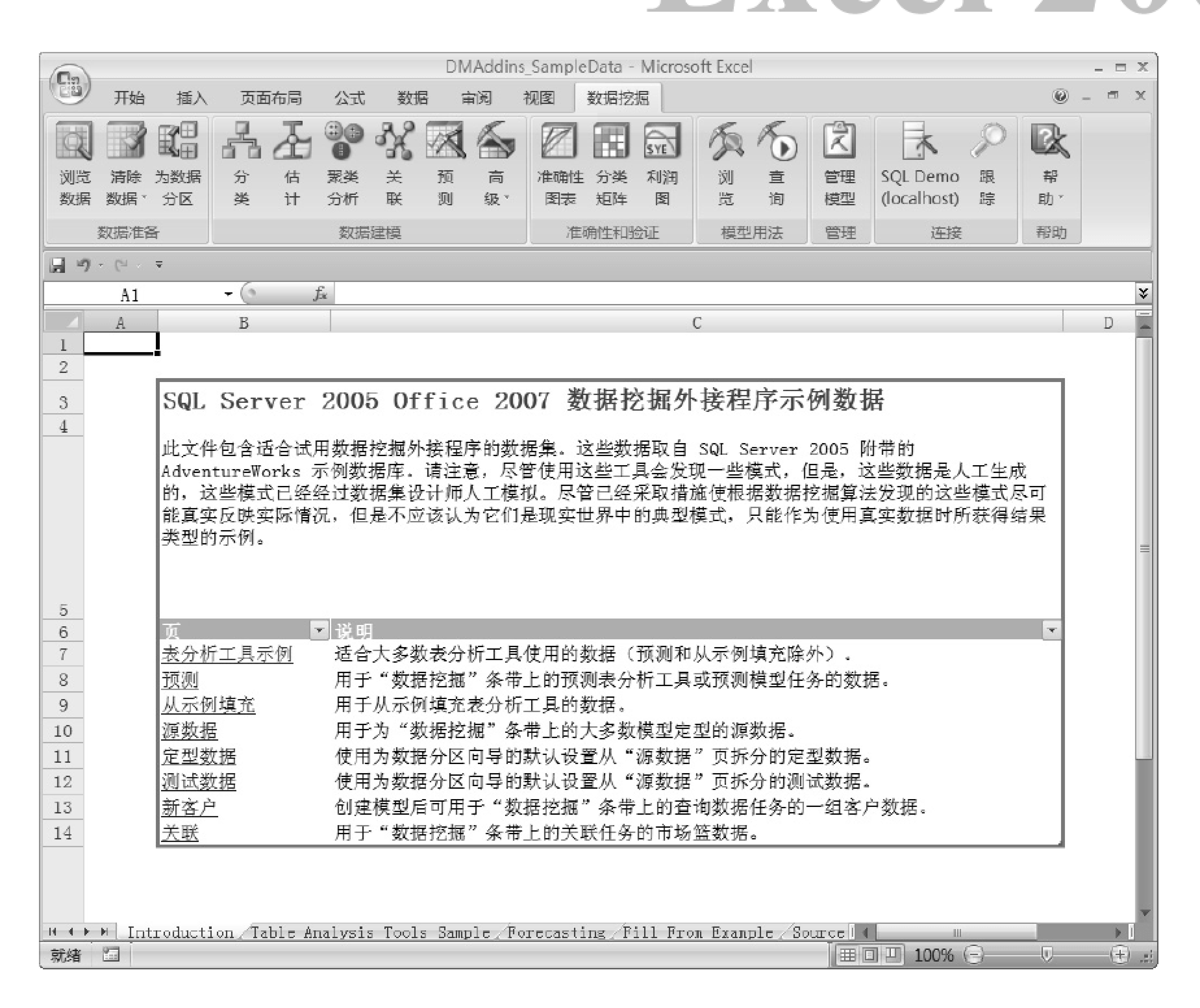
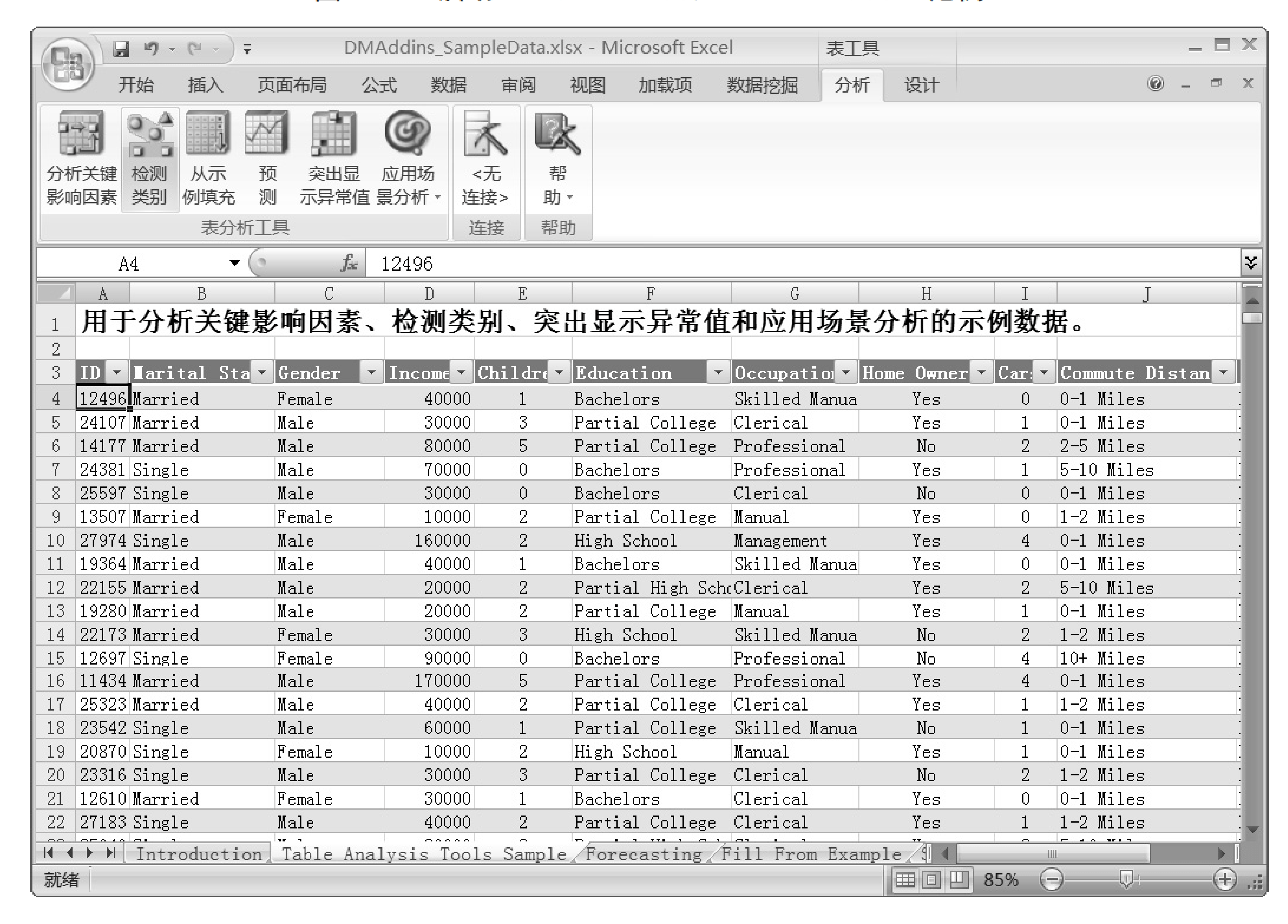


图 17-4　启动 Excel 2007 SQL 2005 DM addin 范例



ID	Marital Sta	Gender	Income	Childr	Education	Occupatio	Home Owner	Car	Commute Distan
12496	Married	Female	40000	1	Bachelors	Skilled Manua	Yes	0	0-1 Miles
24107	Married	Male	30000	3	Partial College	Clerical	Yes	1	0-1 Miles
14177	Married	Male	80000	5	Partial College	Professional	No	2	2-5 Miles
24381	Single	Male	70000	0	Bachelors	Professional	Yes	1	5-10 Miles
25597	Single	Male	30000	0	Bachelors	Clerical	No	0	0-1 Miles
13507	Married	Female	10000	2	Partial College	Manual	Yes	0	1-2 Miles
27974	Single	Male	160000	2	High School	Management	Yes	4	0-1 Miles
19364	Married	Male	40000	1	Bachelors	Skilled Manua	Yes	0	0-1 Miles
22155	Married	Male	20000	2	Partial High Schc	Clerical	Yes	2	5-10 Miles
19280	Married	Male	20000	2	Partial College	Manual	Yes	1	0-1 Miles
22173	Married	Female	30000	3	High School	Skilled Manua	No	2	1-2 Miles
12697	Single	Female	90000	0	Bachelors	Professional	No	4	10+ Miles
11434	Married	Male	170000	5	Partial College	Professional	Yes	4	0-1 Miles
25323	Married	Male	40000	2	Partial College	Clerical	Yes	1	1-2 Miles
23542	Single	Male	60000	1	Partial College	Skilled Manua	No	1	0-1 Miles
20870	Single	Female	10000	2	High School	Manual	Yes	1	0-1 Miles
23316	Single	Male	30000	3	Partial College	Clerical	No	2	1-2 Miles
12610	Married	Female	30000	1	Bachelors	Clerical	Yes	0	0-1 Miles
27183	Single	Male	40000	2	Partial College	Clerical	Yes	1	1-2 Miles

图 17-5　【分析】和【设计】选项

单击【分析关键影响因素】按钮，选择要当作分析目标的单一数据列。也可以单击【选

择分析时要使用的列】超级链接，并选择最可能包含相关数据的数据列，取消选择对于模型分析不重要的数据列，如标识符或名称，单击【确定】按钮。在这里选择 Marital Status 这个变量，如图 17-6 所示。



图 17-6　选择当作分析目标的单一数据列

当利用【分析关键影响因素】工具建立报表时，会执行以下三个作业：

① 建立数据挖掘结构来储存与数据有关的关键信息。

② 使用 Microsoft 贝叶斯概率分类算法来建立数据挖掘模型。

③ 针对指定的每一对属性发出预测查询，以识别能明显区分两个目标属性的因素。此工具在执行数据分析之后会自动设定所有参数，以决定最佳的设定。

建立的报表包含具有下列信息的四个数据列：①数据列包含区别因素的数据列名称；②值与目标之间具有最强关联的值；③偏好此因素所预测的结果或目标值；④表示相对影响程度大小的水平直方图，用以指示关联的强度。

运行完成后，会弹出如图 17-7 所示的【对比报表】对话框，询问是否增加对比报表。

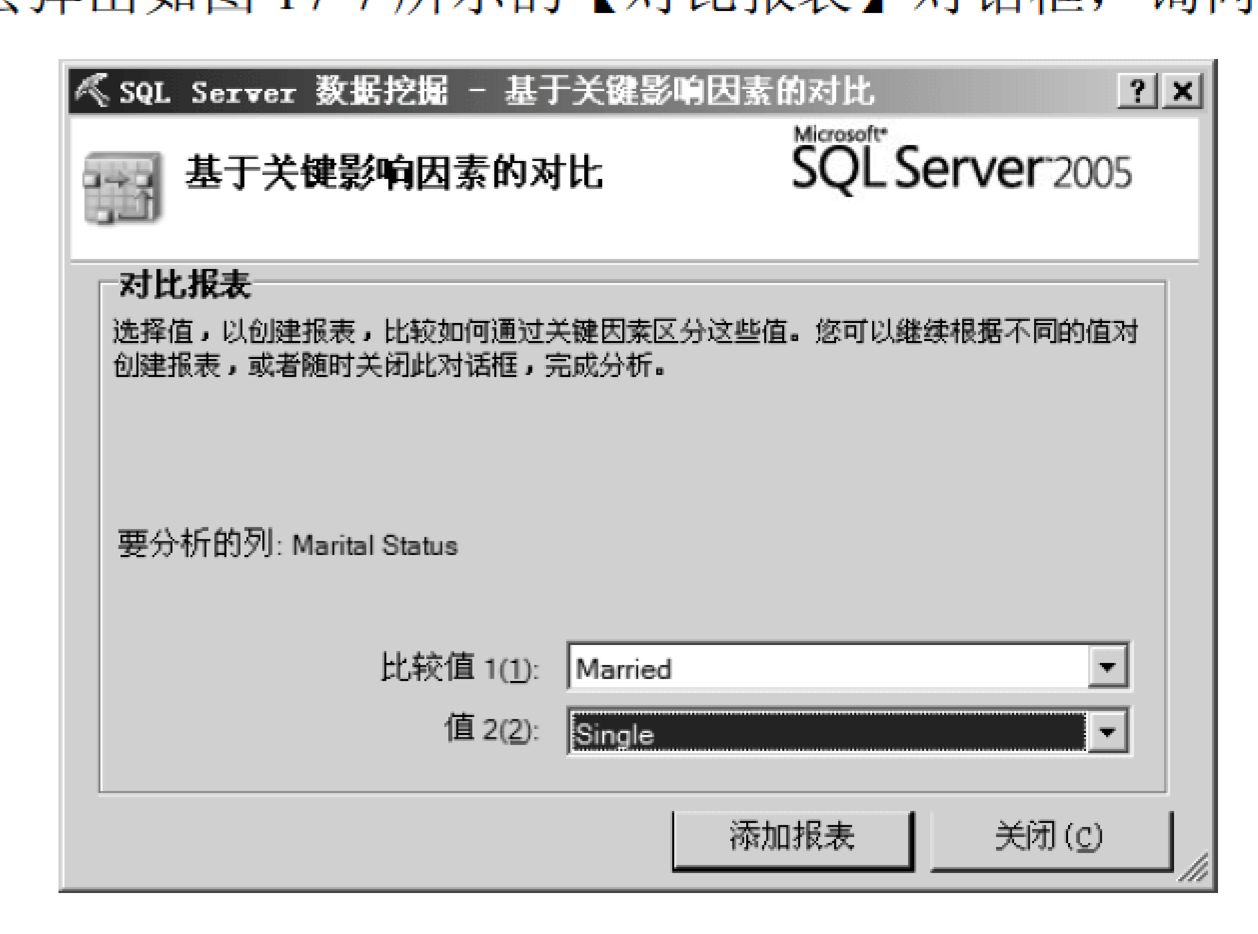
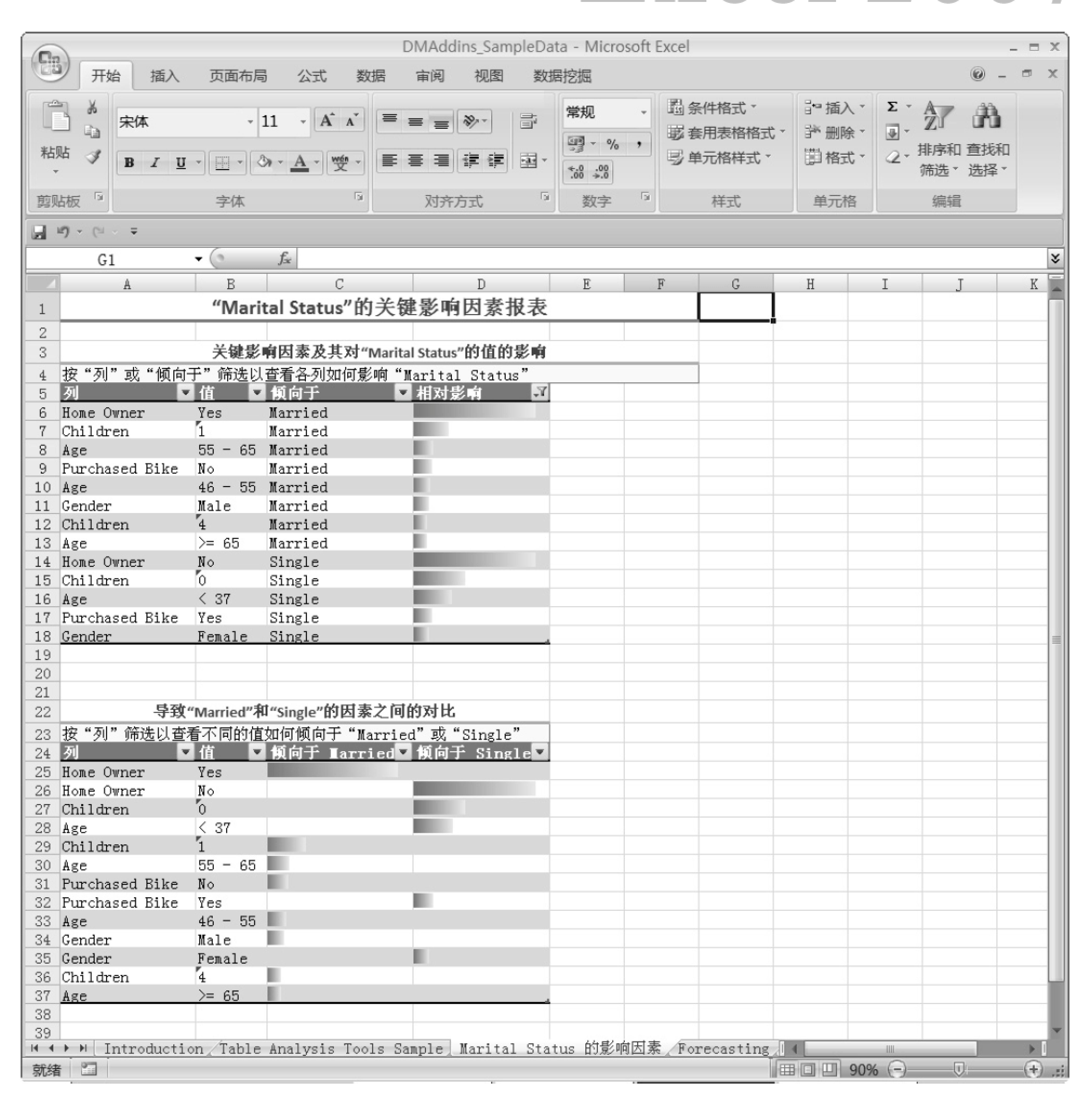


图 17-7　【对比报表】对话框

选择对比值后，单击【确定】按钮，可以得到如图 17-8 所示的报表。



DMAddins_SampleData - Microsoft Excel
开始 插入 页面布局 公式 数据 审阅 视图 数据挖掘
“Marital Status”的关键影响因素报表
关键影响因素及其对“Marital Status”的值的影响
按“列”或“倾向于”筛选以查看各列如何影响“Marital Status”
列 值 倾向于 相对影响
Home Owner Yes Married
Children 1 Married
Age 55 - 65 Married
Purchased Bike No Married
Age 46 - 55 Married
Gender Male Married
Children 4 Married
Age >= 65 Married
Home Owner No Single
Children 0 Single
Age < 37 Single
Purchased Bike Yes Single
Gender Female Single
导致“Married”和“Single”的因素之间的对比
按“列”筛选以查看不同的值如何倾向于“Married”或“Single”
列 值 倾向于 Married 倾向于 Single
Home Owner Yes
Home Owner No
Children 0
Age < 37
Children 1
Age 55 - 65
Purchased Bike No
Purchased Bike Yes
Age 46 - 55
Gender Male
Gender Female
Children 4
Age >= 65
Introduction Table Analysis Tools Sample Marital Status 的影响因素 Forecasting
就绪 90%

图 17-8　分析结果报表

第 18 章　检 测 类 别

在 Excel 2007 SQL 2005 DM addin 分析工具栏下，有一个检测类别目录选项，可以自动检测数据表中所有变量的类别目录。

首先，指定用于分析的数据列。可以不选取有相异值的数据列，例如个人名称或记录标识符，因为这些数据列对分析没有帮助。

其次，选择性地指定要建立的类别目录数目。工具找到多少类别目录，系统就会自动建立同样多的类别目录。单击【运行】按钮，如图 18-1 所示。

工具会建立名为【类别目录报表】的新工作表，工作表中包含类别目录列表及其特性。

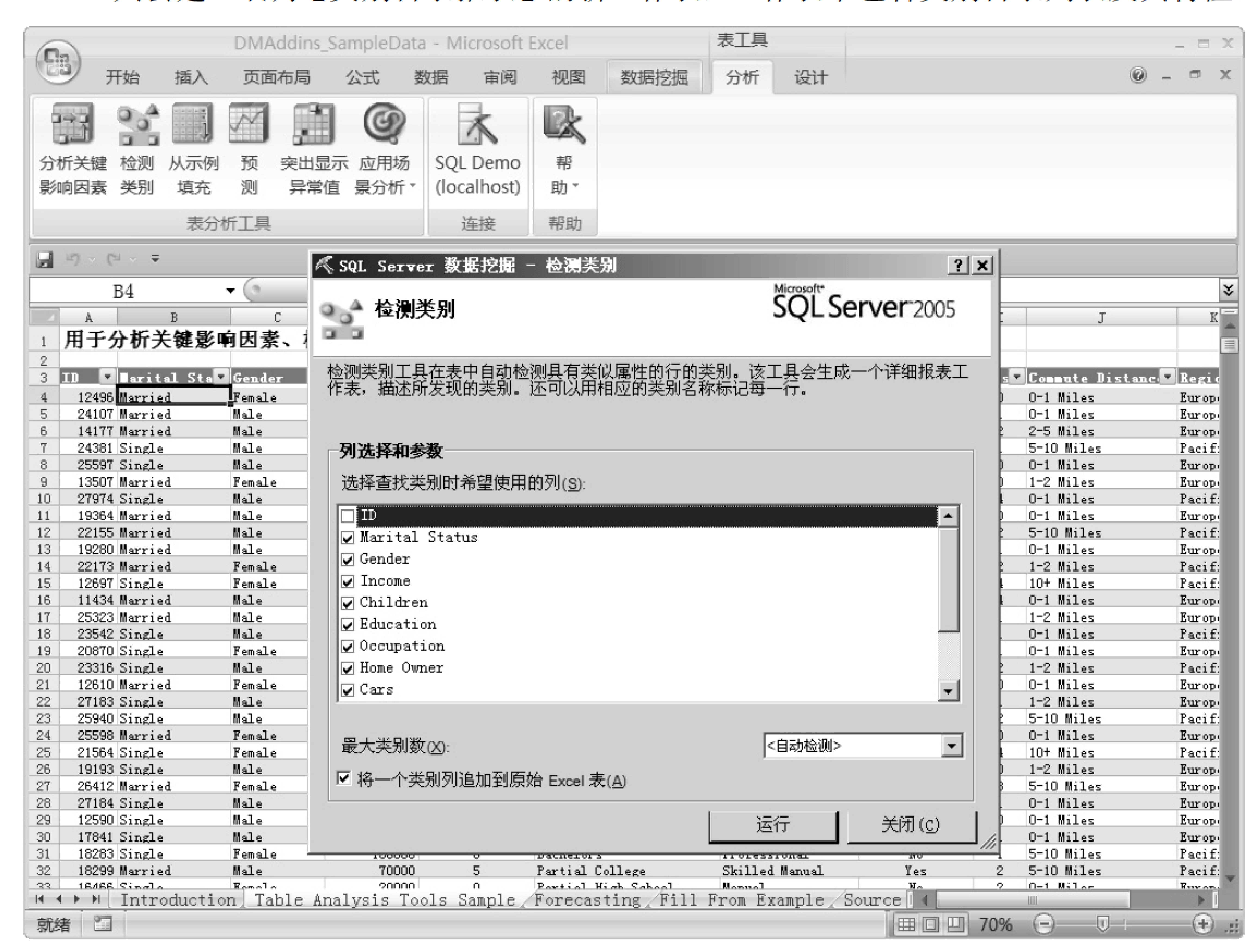


图 18-1　检测类别目录

运行后，会将检测到的类别目录进行统计，并在新的 Excel 窗口中显示，如图 18-2 所示。

第一个数据表会以默认名称如“类别 1”、“类别 2”等来列出新的类别目录。若要让类别目录更容易使用，可以检阅特性列表并且对类别目录指派新名称。例如，如果类别目录 1 的特性包含客户年龄和地区，可以在上方图表的“类别 1”名称上单击，然后输入所要的

类别项目名称。新的类别名称会自动更新。

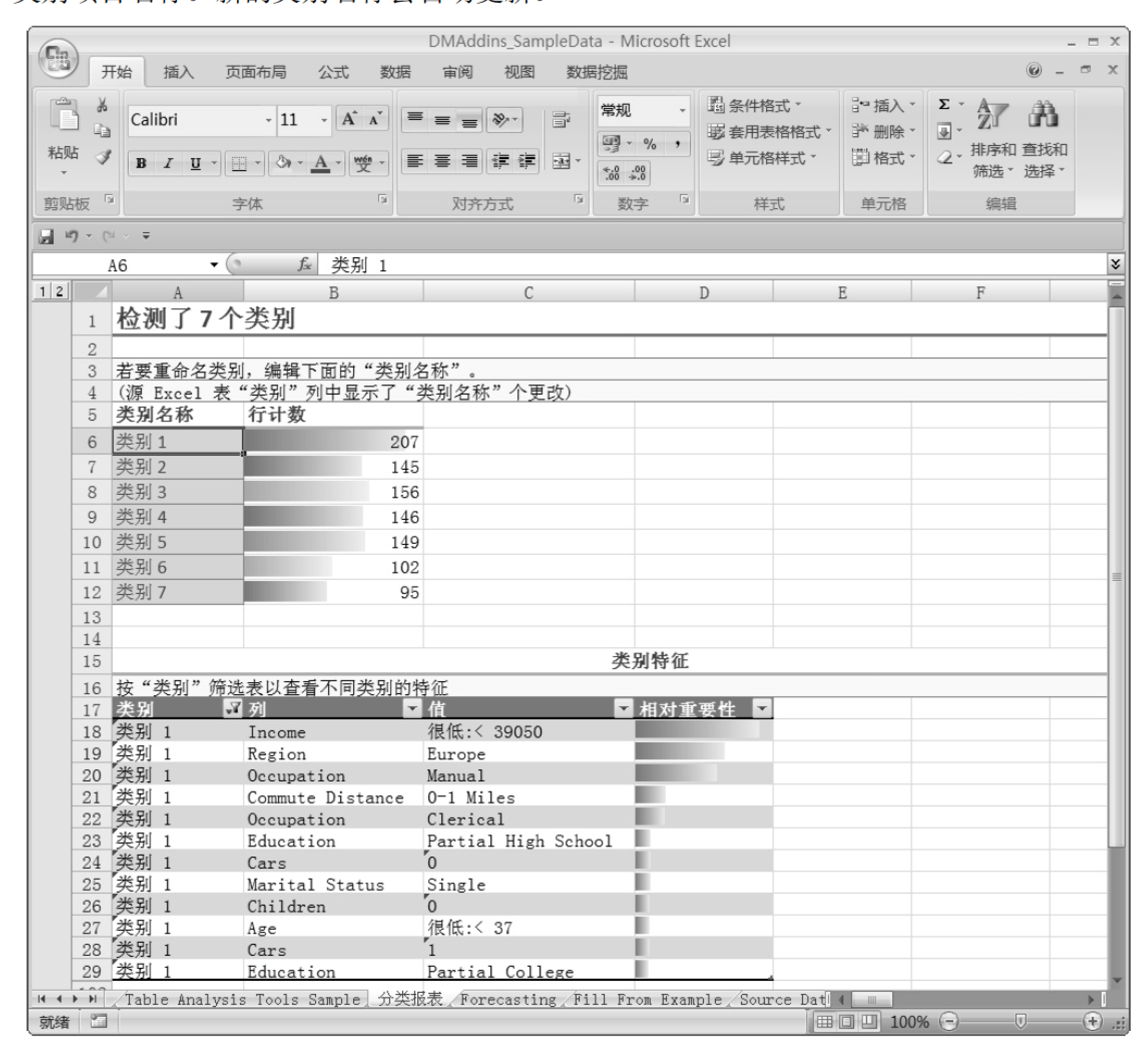


图 18-2　检测到的类别目录

第一个数据表也会显示原始数据中分成该类别目录的数据列数目。

第二个数据表“类别特性”会显示在类别目录中所找到的相似性的详细数据。在类别目录数据列顶端，单击【筛选】按钮即可查看每个类别目录的特性。类别目录的特性包含下列信息如图 18-3 所示。

数据行：数据行名称，一般是属性，例如收入、年龄、教育程度。

相对重要性：颜色条，表示属性和值组作为区别因素的重要性。颜色条越长，此属性代表此类别目录的可能性就越大。

在类别目录报表底部单击图表时，Excel 会显示【数据透视表字段列表】图表控件，可以以互动方式筛选及重新排列字段，如图 18-4 所示。



图 18-3 类别特性

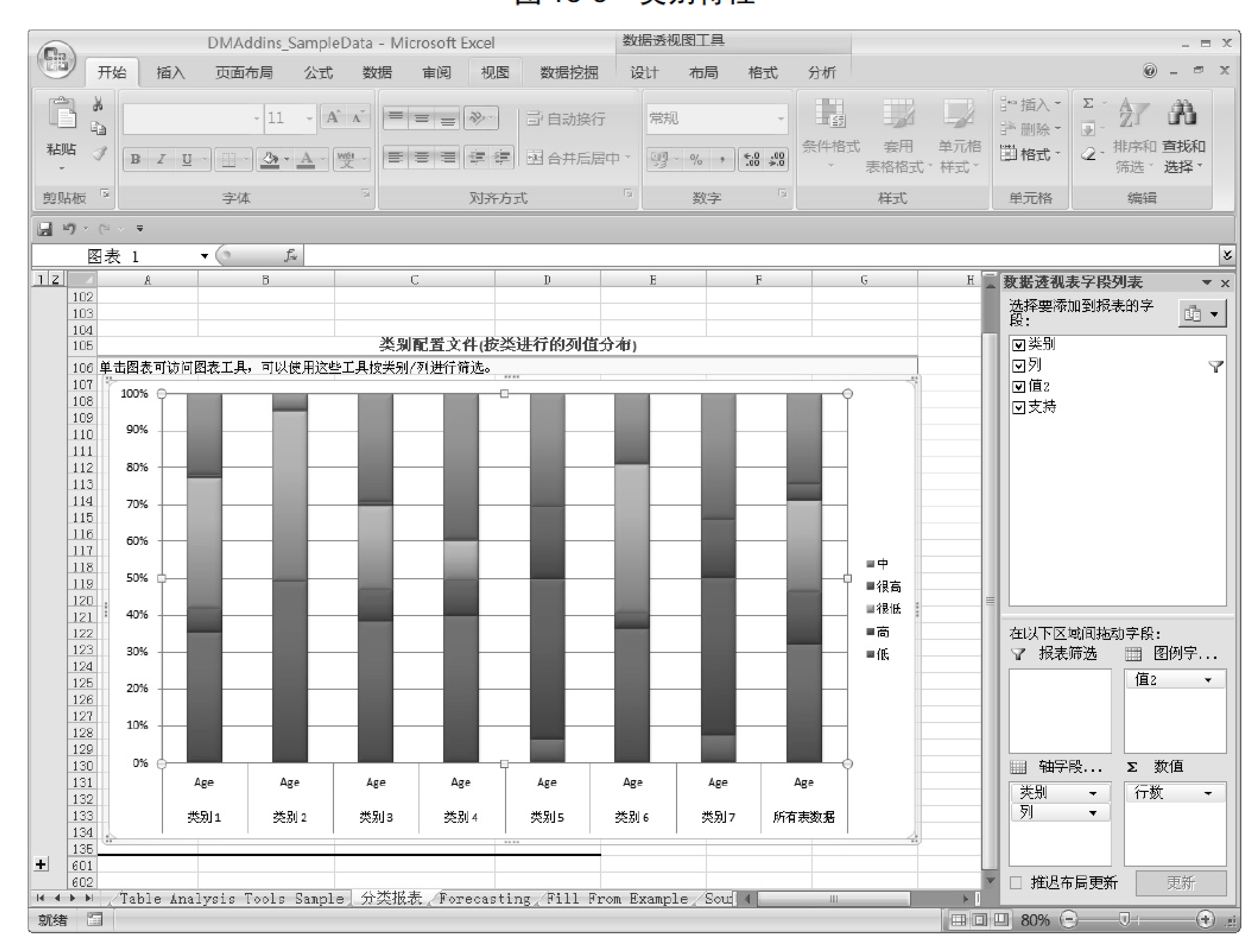


图 18-4 【数据透视表字段列表】图表控件

第 19 章　从示例填充

首先启动 Excel 2007 SQL 2005 DM addin 的从示例填充范例。从示例填充工具实现在 Excel 数据表快速建立新的数据行，同时若用户建立新值，该工具也可以通过分析范例模型来填充新值的数据列。例如，在列出客户及其年度购买额的数据表中，可以建立几种范例的新数据列和类型，如 High Value Customer 和其他等，如图 19-1 所示。此工具会分析数据中的现有模型，并会套用已输入的范例，填满数据列其余部分的值。如果用户对结果不满意，即可提供更多范例以改善结果。



图 19-1　从示例填充

也可以指定可能对预测遗失数据值最有帮助的数据列，以便自定义结果。例如，如果从经验得知，在一个数据列和一个具有遗失值的数据列之间有较强的关系，即可取消选取其他数据列以取得较佳结果。

分析完成后会建立包含分析结果的新工作表。名为“<数据列名称>模型”的新工作表，会报告找到的数据列规则（或称关键影响因素），并显示每条规则的概率。

如果向导检测到模式，便会将包含新值的数据列新增到原始的数据表。可以检阅这些值，并且将其与原始的值比较。

模式报表会显示所预测的每个值的关键影响因素。每个影响因素或规则都会被描述为数据列、该数据列中的值，以及规则对于预测的相对影响的组合，如图 19-2 所示。

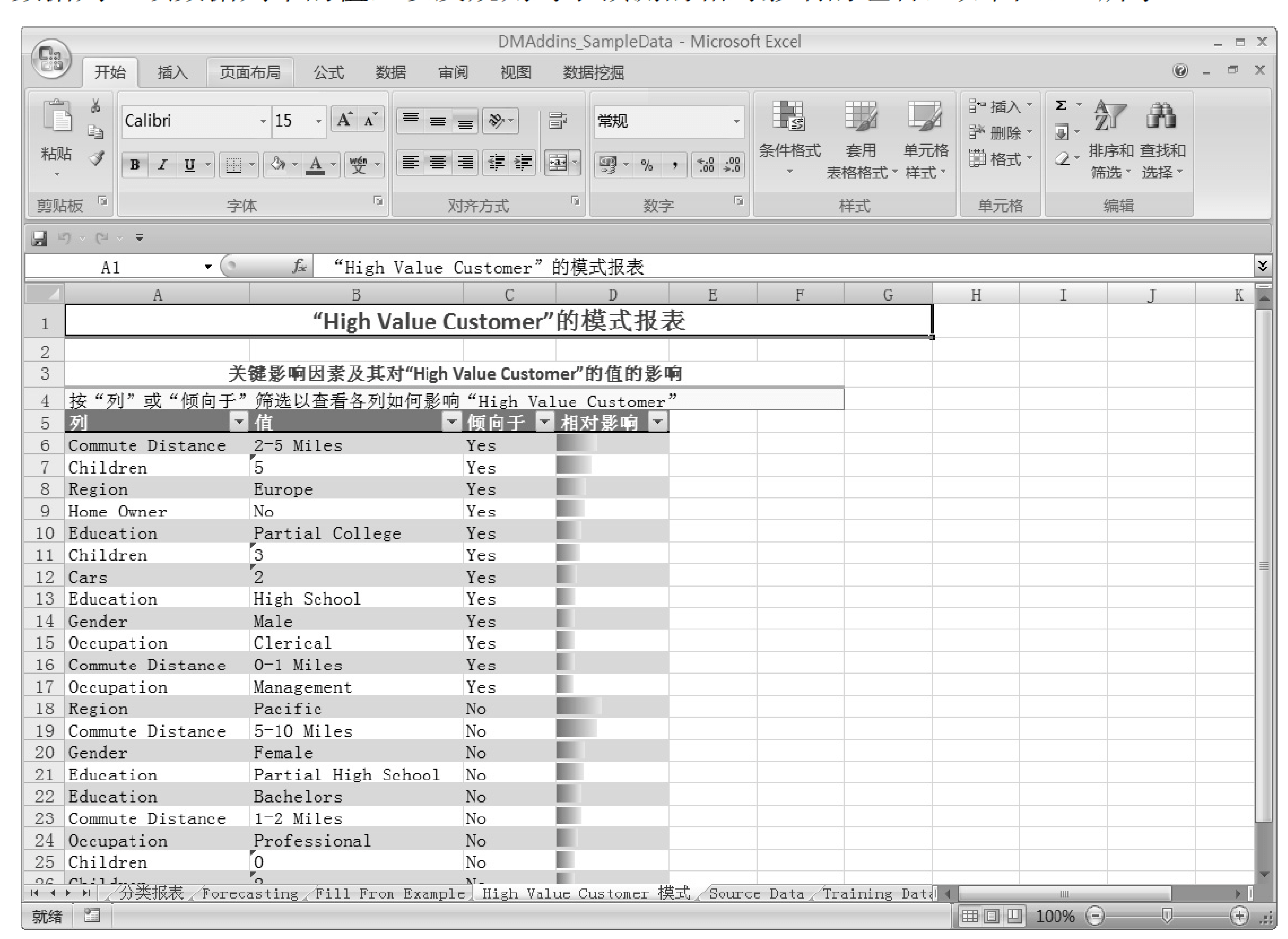


图 19-2　模式报表

第 20 章　预　　测

启动 Excel 2007 SQL 2005 DM addin 的预测范例，该范例为三个不同地区 2001 年 7 月至 2004 年 6 月 M200 型号的销售记录。预测工具可根据 Excel 数据表或其他数据来源中的数据进行预测，并且可以选择性地查看与每个预测值相关的概率，如图 20-1 所示。例如，如果数据包含日期和当月每日总销售额的数据列，可以预测未来的销售情况，也可以指定要预测的数目，例如，可以预测 10 天或 20 天。

向导完成后，会将新的预测值附加到来源数据表末尾，并且突出显示。但新的时间序列值不会附加，可以先检阅预测。

向导也会建立名为“预测报表”的新工作表。这个工作表会报告向导是否成功建立预测。新工作表也包含显示历史趋势的折线图。

若将新预测值加入到原来的时间序列数据列后，预测值会加入折线图。原有记录值用实线表示，预测值则用虚线表示。

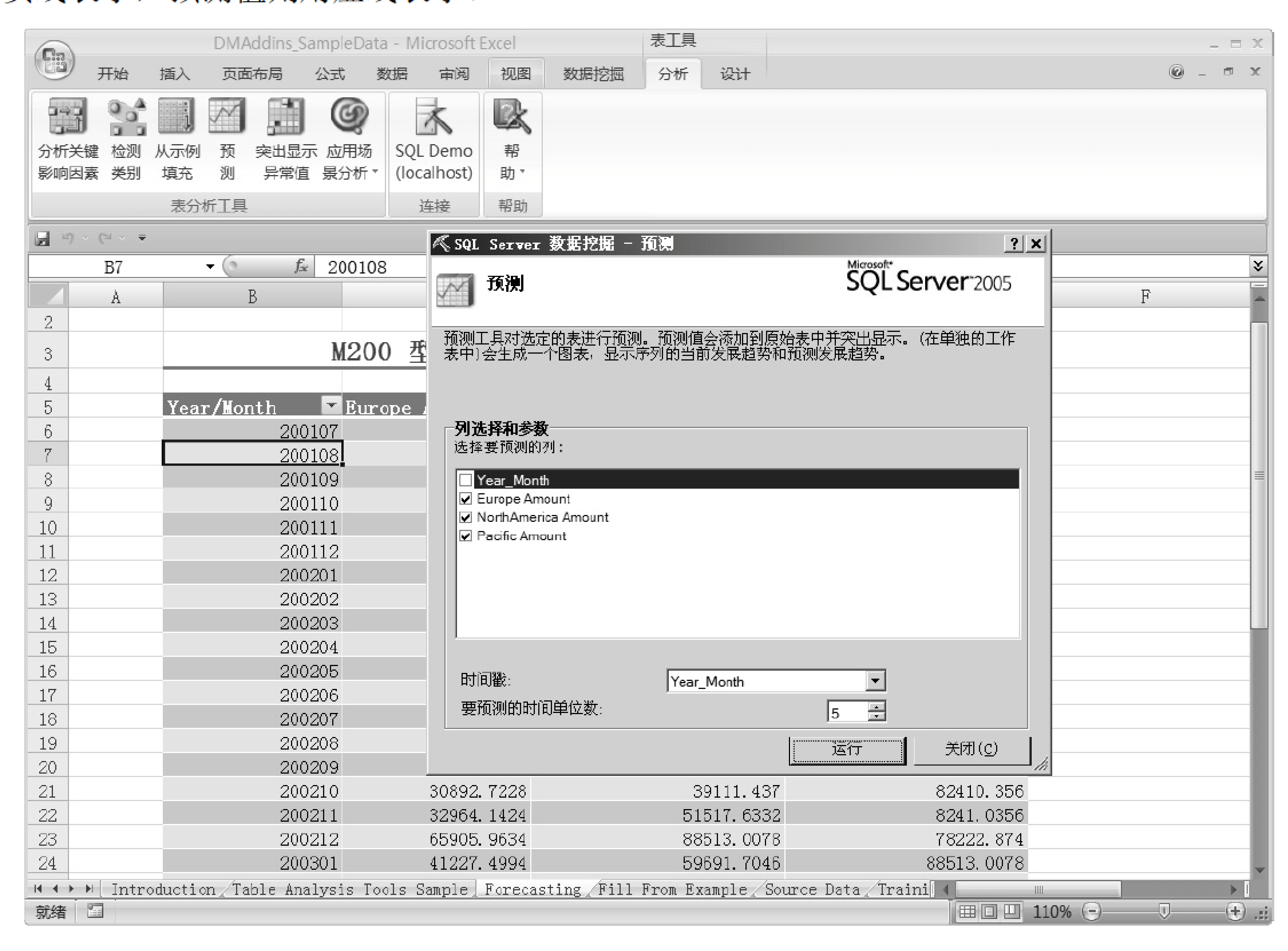
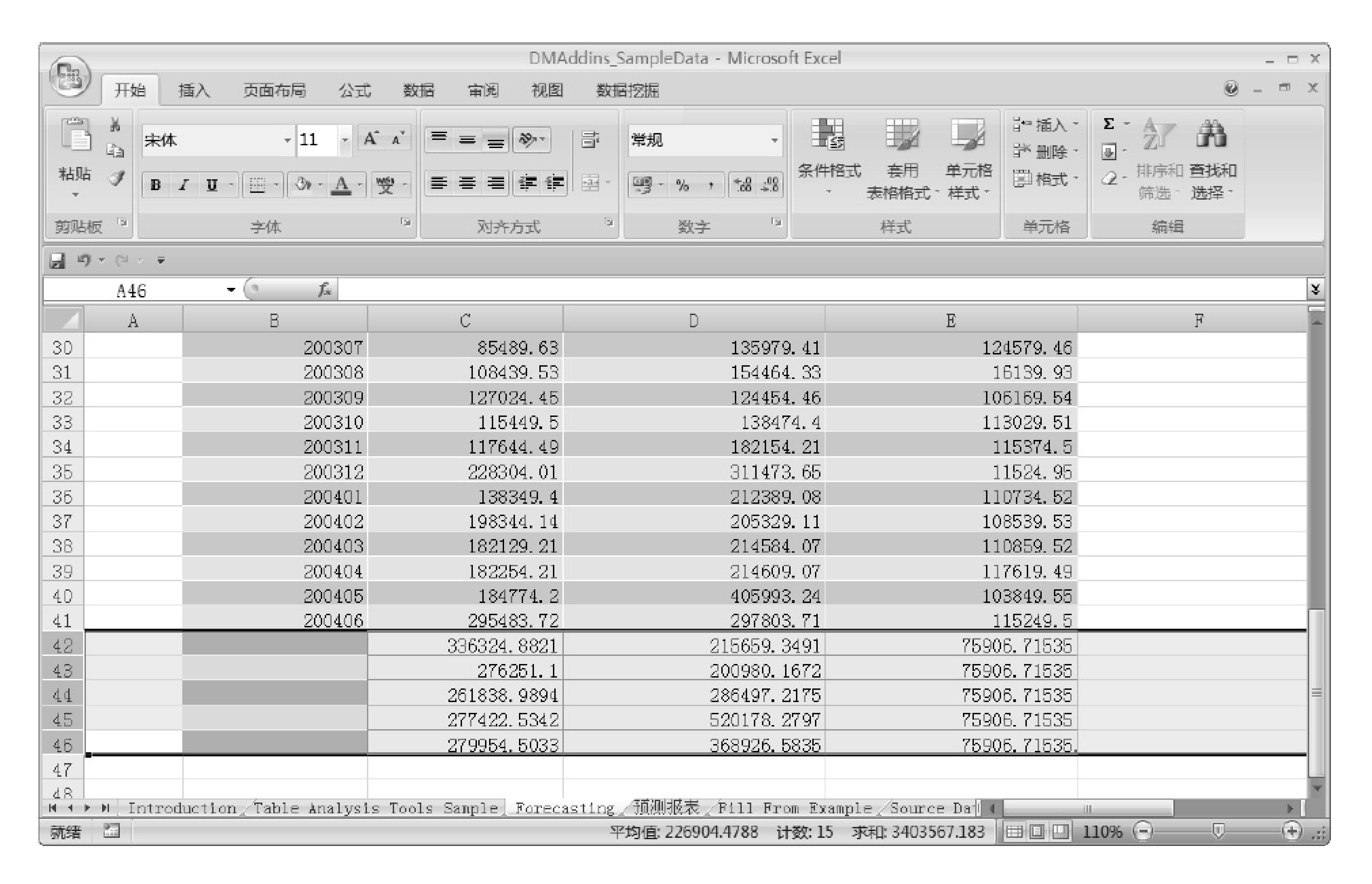


图 20-1　预测

图 20-2 是附加在原时序数据后的预测值，共有 5 期。



	A	B	C	D	E
30		200307	85489.63	135979.41	124579.46
31		200308	108439.53	154464.33	16139.93
32		200309	127024.45	124454.46	106169.54
33		200310	115449.5	138474.4	113029.51
34		200311	117644.49	182154.21	115374.5
35		200312	228304.01	311473.65	11524.95
36		200401	138349.4	212389.08	110734.52
37		200402	198344.14	205329.11	108539.53
38		200403	182129.21	214584.07	110859.52
39		200404	182254.21	214609.07	117619.49
40		200405	184774.2	405993.24	103849.55
41		200406	295483.72	297803.71	115249.5
42			336324.8821	215659.3491	75906.71535
43			276251.1	200980.1672	75906.71535
44			261838.9894	286497.2175	75906.71535
45			277422.5342	520178.2797	75906.71535
46			279954.5033	368926.5835	75906.71535

图 20-2　预测值

三个地区未来 5 个月的销售记录的预测折线图如图 20-3 所示。

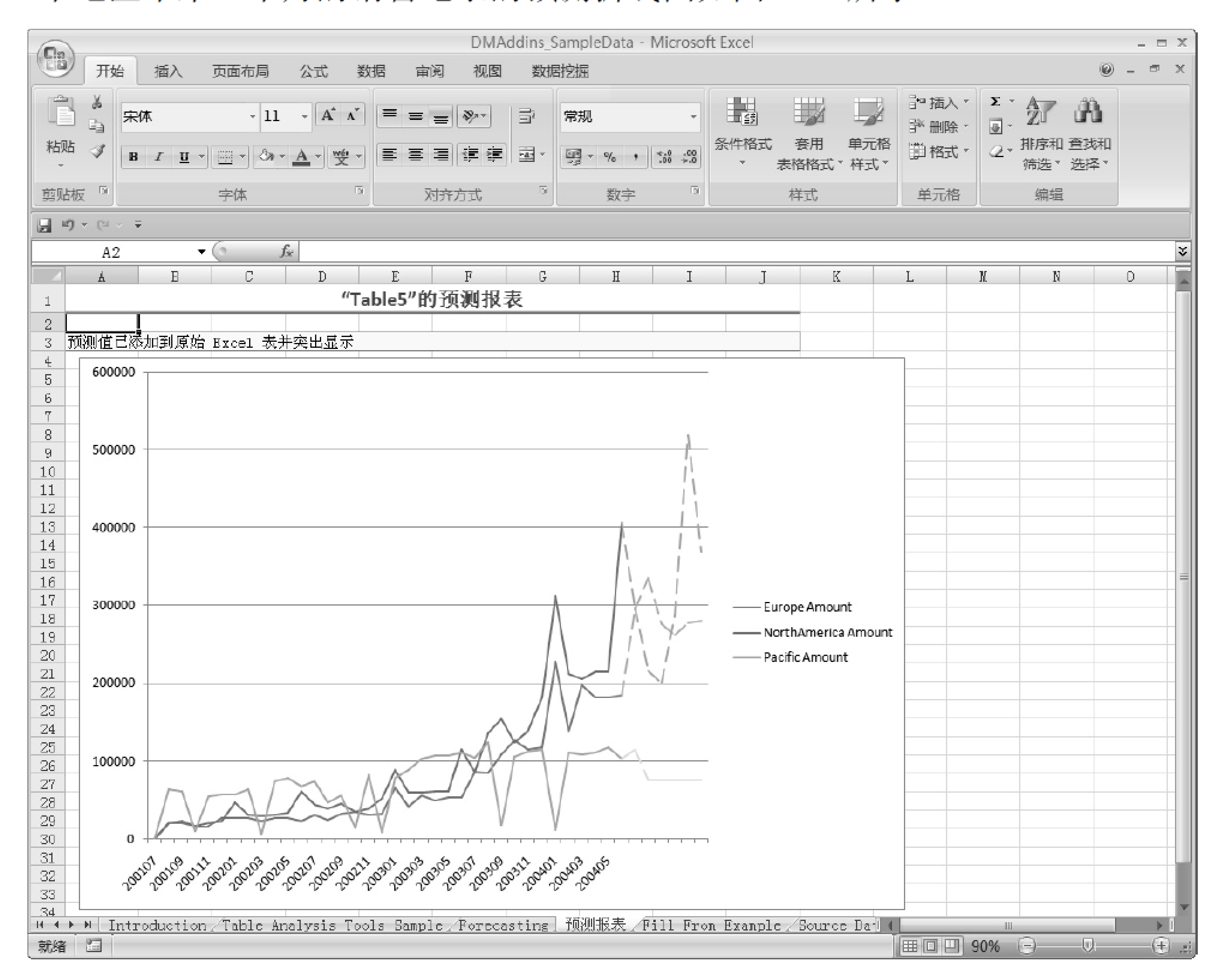


图 20-3　预测折线图

第 21 章　突出显示异常值

首先开启 Excel 2007 SQL 2005 DM addin 的 Table Analysis Tools Sample 范例数据表。因为数据输入错误或反映不寻常趋势的错误的值通称为异常值（Outlier），例如家长的年龄为 3 岁，显然是预期值以外的值。但不论是哪一个情况，异常状况都可能影响到分析的质量。因此，突出显示异常值工具有助于用户寻找到这些值，逐个查看，并进一步找出原因等以提高分析的质量，如图 21-1 所示。

突出显示异常值工具适用于 Excel 数据表中的整个数据范围，也可以只选取几个数据列，并可以调整控制数据变化的临界值，以寻找更多或更少的例外状况。

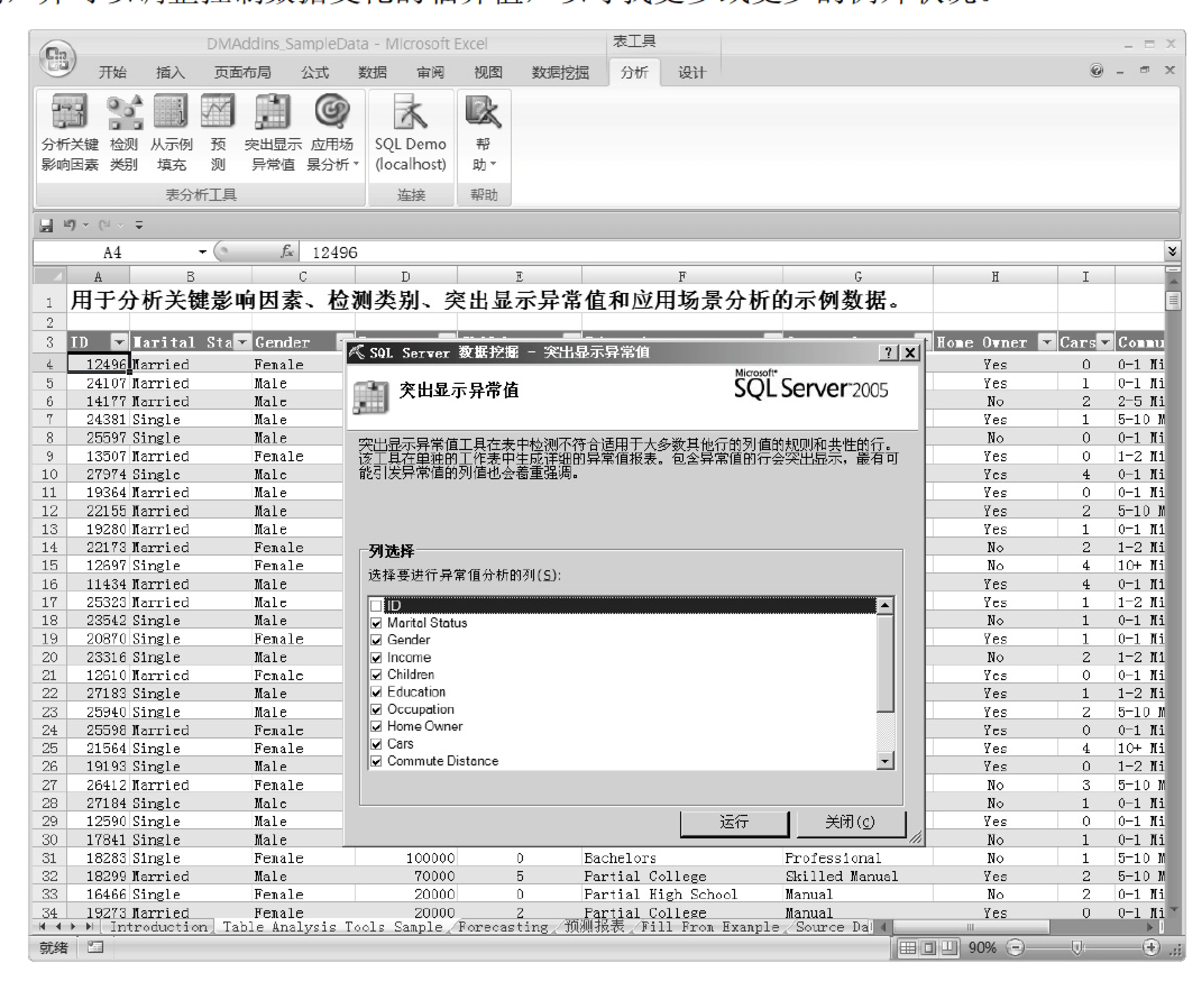
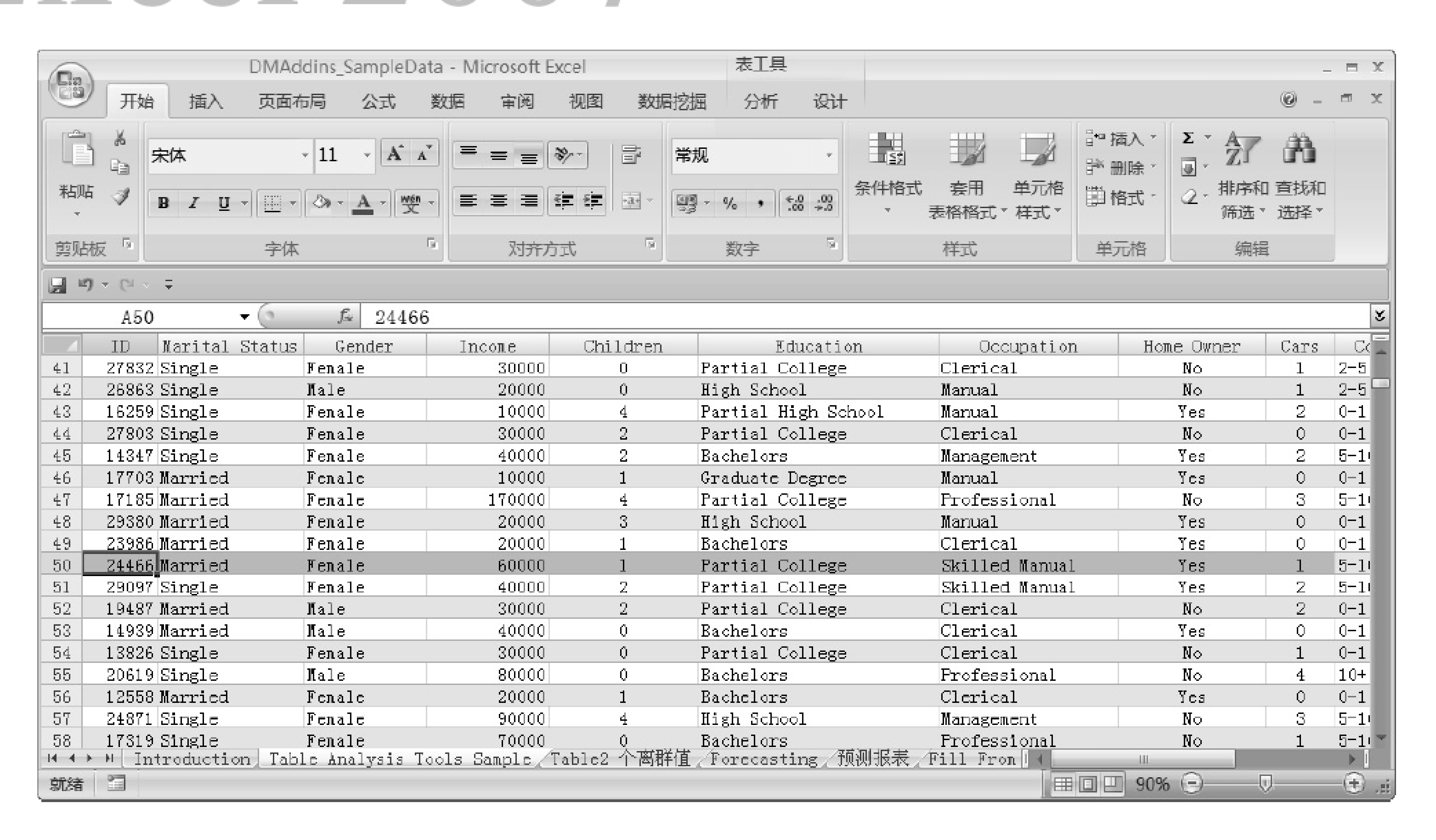


图 21-1　突出显示异常值

当向导完成时，会建立一张新的工作表，其中包含分析的每一个数据列中找到的异常值数目的摘要报表，如图 21-2 所示。此工具也会在原始的数据表中突出显示异常值。



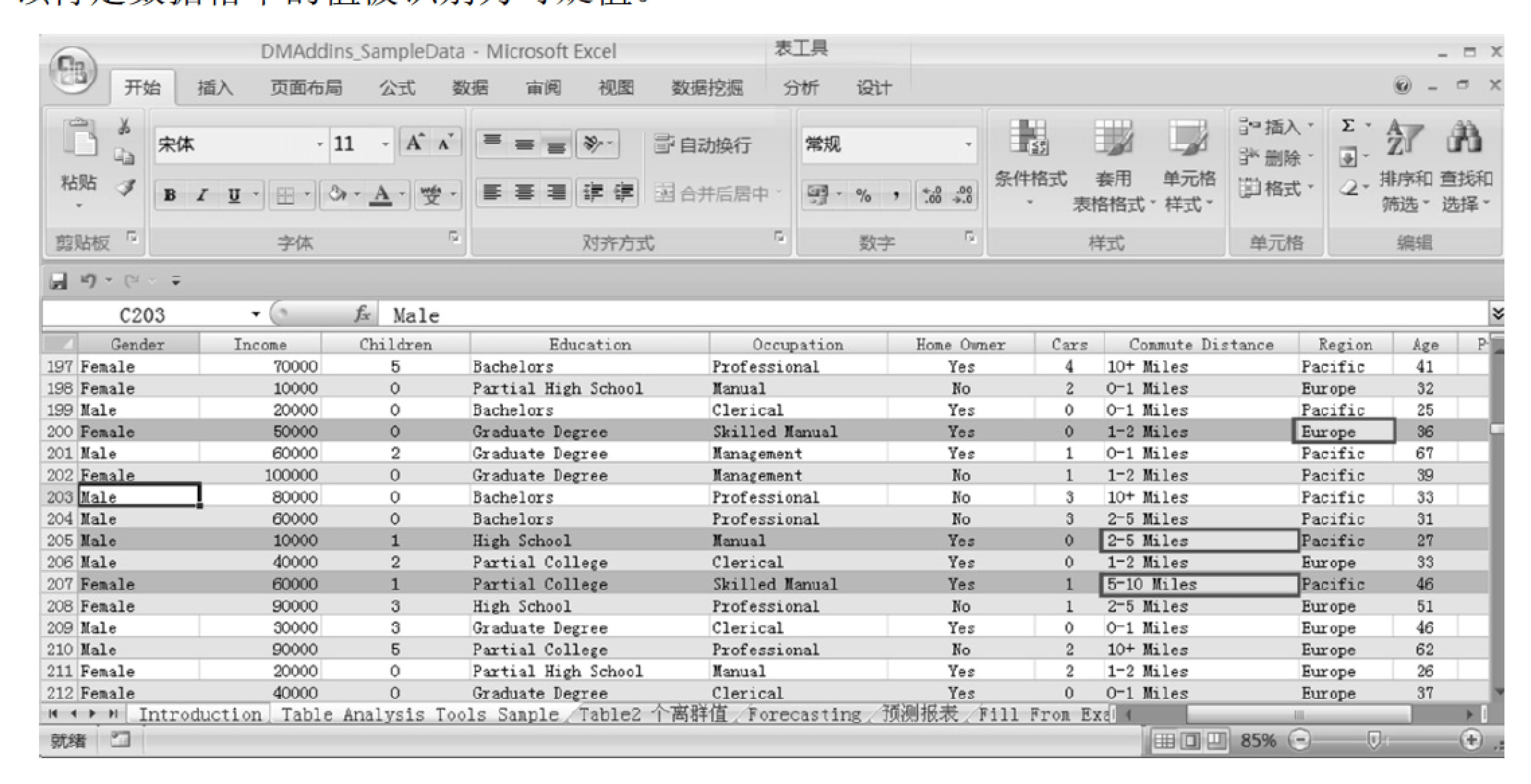
A50 | 24466

	ID	Marital Status	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Co
41	27832	Single	Female	30000	0	Partial College	Clerical	No	1	2-5
42	26863	Single	Male	20000	0	High School	Manual	No	1	2-5
43	16259	Single	Female	10000	4	Partial High School	Manual	Yes	2	0-1
44	27803	Single	Female	30000	2	Partial College	Clerical	No	0	0-1
45	14347	Single	Female	40000	2	Bachelors	Management	Yes	2	5-1
46	17703	Married	Female	10000	1	Graduate Degree	Manual	Yes	0	0-1
47	17185	Married	Female	170000	4	Partial College	Professional	No	3	5-1
48	29380	Married	Female	20000	3	High School	Manual	Yes	0	0-1
49	23986	Married	Female	20000	1	Bachelors	Clerical	Yes	0	0-1
50	24466	Married	Female	60000	1	Partial College	Skilled Manual	Yes	1	5-1
51	29097	Single	Female	40000	2	Partial College	Skilled Manual	Yes	2	5-1
52	19487	Married	Male	30000	2	Partial College	Clerical	No	2	0-1
53	14939	Married	Male	40000	0	Bachelors	Clerical	Yes	0	0-1
54	13826	Single	Female	30000	0	Partial College	Clerical	No	1	0-1
55	20619	Single	Male	80000	0	Bachelors	Professional	No	4	10+
56	12558	Married	Female	20000	1	Bachelors	Clerical	Yes	0	0-1
57	24871	Single	Female	90000	4	High School	Management	No	3	5-1
58	17319	Single	Female	70000	0	Bachelors	Professional	No	1	5-1

图 21-2　摘要报表

由于突出显示异常值工具会分析整体趋势，所以它可能会发现数据列中的大多数值为正常值，而只突出显示该数据列中的一个数据格。

突出显示异常值工具会突出显示原始数据表中可疑的数据格，如图 21-3 所示。如果突出显示的颜色很深，则表示这一数据行需要特别留意；如果突出显示的颜色很亮，则表示该特定数据格中的值被识别为可疑值。



C203 | Male

	Gender	Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age
197	Female	70000	5	Bachelors	Professional	Yes	4	10+ Miles	Pacific	41
198	Female	10000	0	Partial High School	Manual	No	2	0-1 Miles	Europe	32
199	Male	20000	0	Bachelors	Clerical	Yes	0	0-1 Miles	Pacific	25
200	Female	50000	0	Graduate Degree	Skilled Manual	Yes	0	1-2 Miles	Europe	36
201	Male	60000	2	Graduate Degree	Management	Yes	1	0-1 Miles	Pacific	67
202	Female	100000	0	Graduate Degree	Management	No	1	1-2 Miles	Pacific	39
203	Male	80000	0	Bachelors	Professional	No	3	10+ Miles	Pacific	33
204	Male	60000	0	Bachelors	Professional	No	3	2-5 Miles	Pacific	31
205	Male	10000	1	High School	Manual	Yes	0	2-5 Miles	Pacific	27
206	Male	40000	2	Partial College	Clerical	Yes	0	1-2 Miles	Europe	33
207	Female	60000	1	Partial College	Skilled Manual	Yes	1	5-10 Miles	Pacific	46
208	Female	90000	3	High School	Professional	No	1	2-5 Miles	Europe	51
209	Male	30000	3	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	46
210	Male	90000	5	Partial College	Professional	No	2	10+ Miles	Europe	62
211	Female	20000	0	Partial High School	Manual	Yes	2	1-2 Miles	Europe	26
212	Female	40000	0	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	37

图 21-3　突出显示可疑的数据格

在查看突出显示的数据格后，可以回到摘要报表，并更改异常阈值，如图 21-4 所示。它表示特定数据格包含异常值的概率，当增加这个值时，它会筛选掉概率较低的值；相反，当减小这个值时，将会看到更多突出显示的数据格。摘要图表会显示每一个数据列中在例外状况临界值以上的数据格数目。

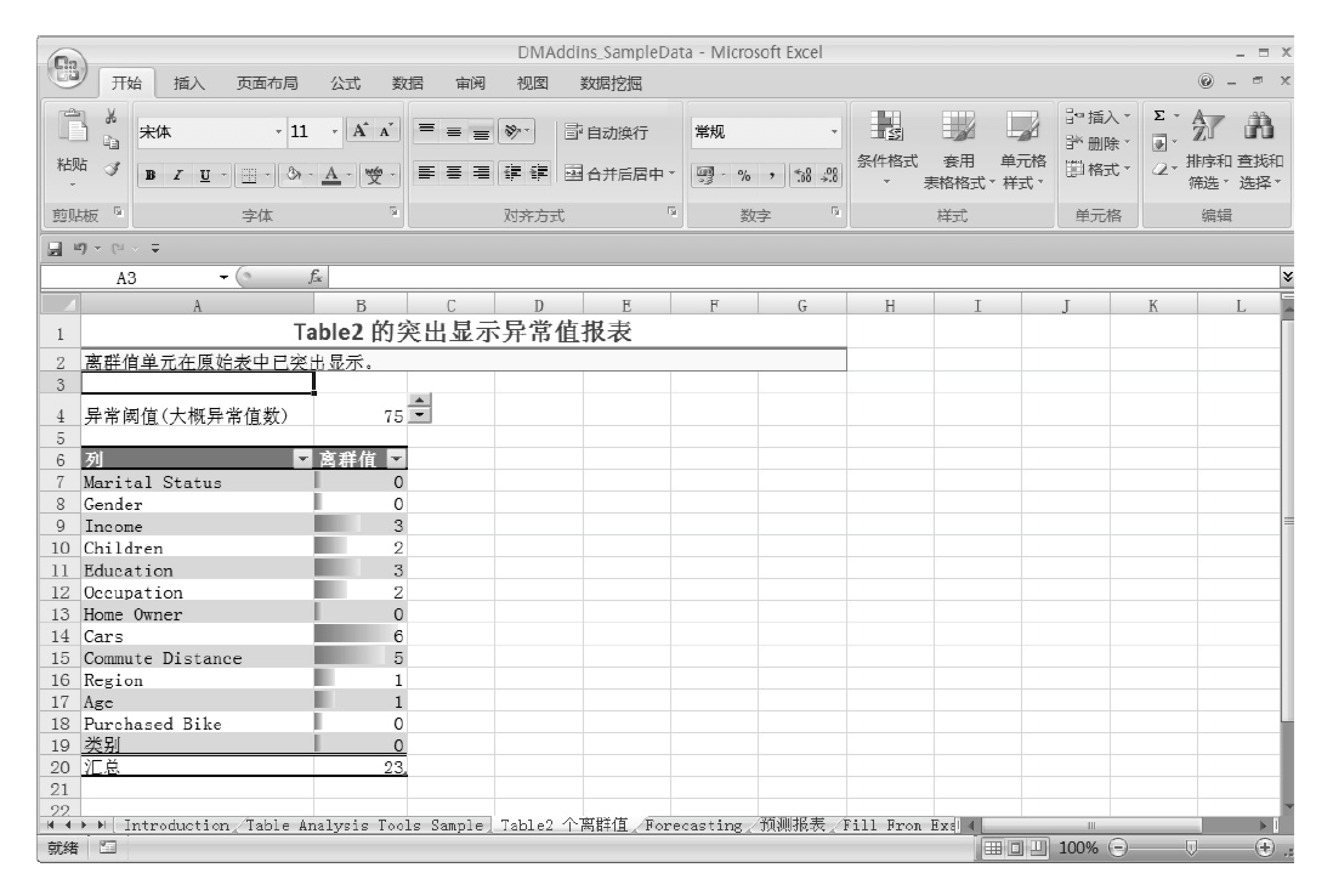


图 21-4　更改异常阈值

当单击【运行】按钮时，此工具会执行以下三个作业：

① 根据目前在数据表中的数据建立数据挖掘结构。

② 使用 Microsoft 聚类算法建立新的数据挖掘模型。

③ 根据模型建立预测查询，以判断工作表中是否存在异常值。

第 22 章　应用场景分析

22.1　目 标 查 找

应用场景分析有两个很有用的工具：目标查找工具和假设工具，如图 22-1 所示。目标查找工具与假设工具相辅相成，假设工具会显示变化的影响，而目标查找工具则会显示为了实现上述变化而必须变化哪些基础因素。

当此工具完成分析时，它会在来源数据表中建立两个新的数据列。这些数据列会显示预测的成功及建议的变化 （如果有的话）。目标查找的操作步骤如下：

图 22-1　目标查找和假设工具

Step1：单击【应用场景分析】按钮，然后单击【目标查找】按钮。

Step2：在【目标查找】对话框中（图 22-2），选择包含目标值的数据列，在这里选择“教育程度”。

Step3：指定要查找的精确值，或是想要增加或减少的值的部分。如果数据列目标包含连续的数值，也可以指定某个范围当作目标。

Step4：指定要变化的数据列。不需要指定数据列的变化数量，系统将会自动评估所有可能的变化值。

Step5：可以选择性地单击【选择分析时要使用的列】超级链接，并选择包含有用信息的数据列，取消选择对于分析没有用处的数据列。

Step6：指定要针对整份数据表还是只有选中的数据列做出预测。

Step7：如果选中【整个表】单选按钮，此工具会将预测加入到来源数据表的两个新数据列中。

Step8：如果选中【当前行】单选按钮，分析的结果会输出到对话框。此对话框会维持可用的状态，可以继续输入新的目标。

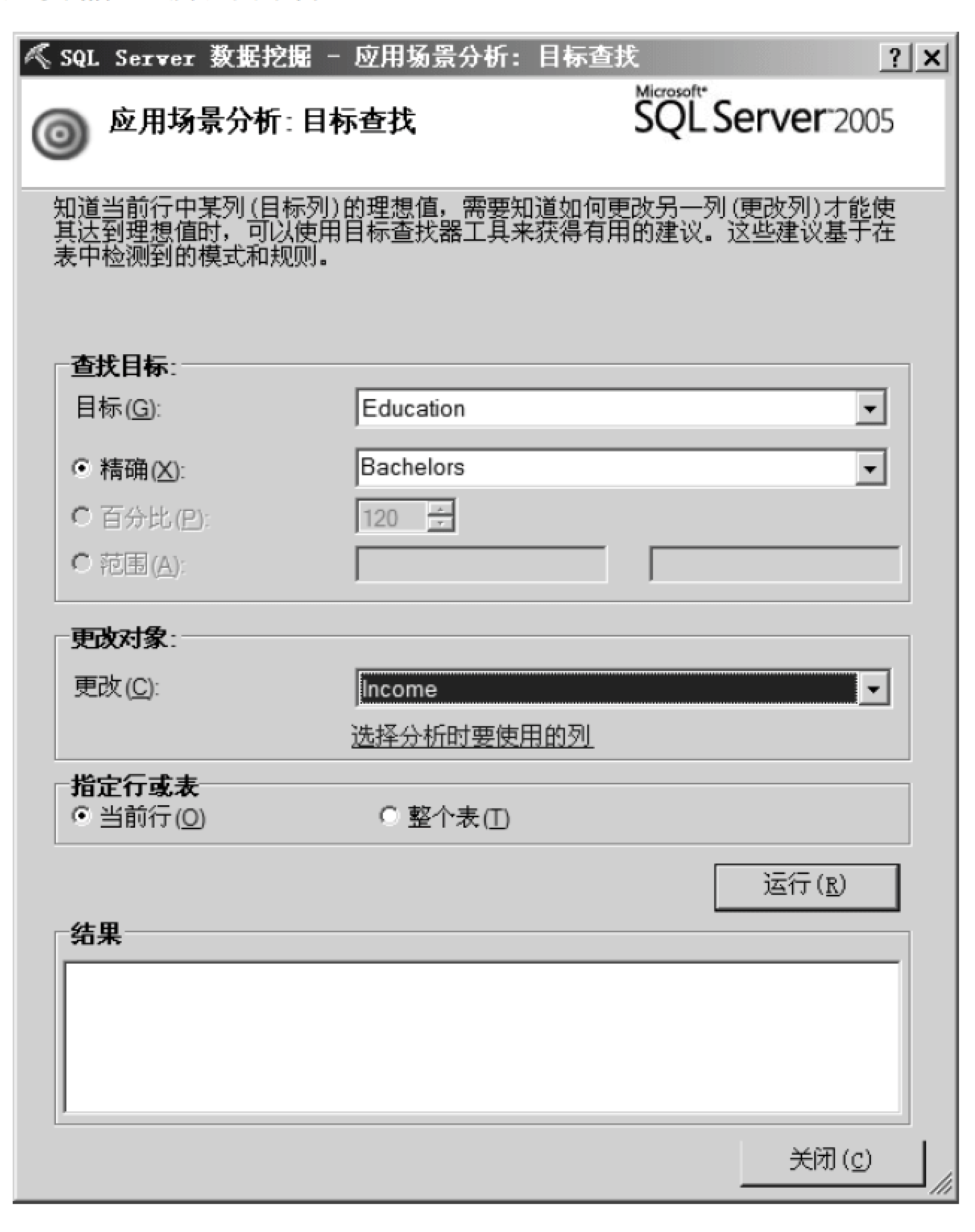
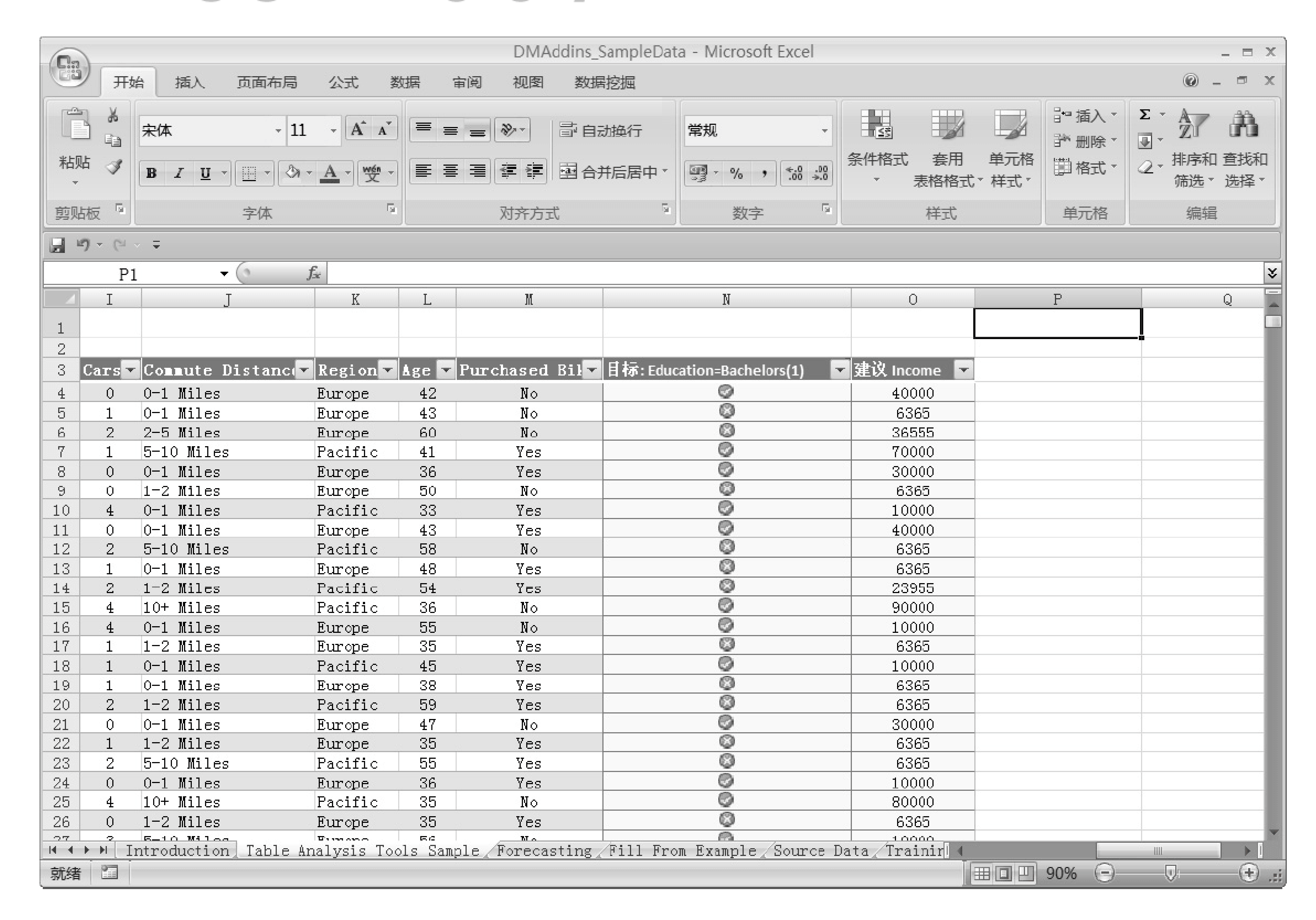


图 22-2 【目标查找】对话框

如图 22-3 所示为针对整份数据表。



Cars	Commute Distance	Region	Age	Purchased Bike	目标: Education=Bachelors(1)	建议 Income
0	0-1 Miles	Europe	42	No		40000
1	0-1 Miles	Europe	43	No		6365
2	2-5 Miles	Europe	60	No		36555
1	5-10 Miles	Pacific	41	Yes		70000
0	0-1 Miles	Europe	36	Yes		30000
0	1-2 Miles	Europe	50	No		6365
4	0-1 Miles	Pacific	33	Yes		10000
0	0-1 Miles	Europe	43	Yes		40000
2	5-10 Miles	Pacific	58	No		6365
1	0-1 Miles	Europe	48	Yes		6365
2	1-2 Miles	Pacific	54	Yes		23955
4	10+ Miles	Pacific	36	No		90000
4	0-1 Miles	Europe	55	No		10000
1	1-2 Miles	Europe	35	Yes		6365
1	0-1 Miles	Pacific	45	Yes		10000
1	0-1 Miles	Europe	38	Yes		6365
2	1-2 Miles	Pacific	59	Yes		6365
0	0-1 Miles	Europe	47	No		30000
1	1-2 Miles	Europe	35	Yes		6365
2	5-10 Miles	Pacific	55	Yes		6365
0	0-1 Miles	Europe	36	Yes		10000
4	10+ Miles	Pacific	35	No		80000
0	1-2 Miles	Europe	35	Yes		6365

图 22-3　针对整份数据表

22.2　假设

假设工具会分析现有数据中的模型，然后评估一个数据列的变化对另一个不同数据列形成的效果。例如，可以浏览涨价对总销售额的影响效果，如图 22-4 所示。向导在决定预测数目上十分灵活，同时在完成初始的分析之后，还会让操作者选择是否要预测数据表中所有数据的结果，或者是否要一次输入一组测试值。其操作步骤如下：

Step1：在【假设】对话框中（图 22-5），选择包含所要变化的数据列，并将变化值指定为特定的值或目前值的百分比（增加或减少）。

Step2：在【目标】下拉列表框中，指定要评估其效果的数据列。

Step3：选择性地单击【选择分析时要使用的列】超级链接，选择在进行预测时可能有用的数据列。也可以取消选择在检测模型时可能不太有用的数据列，例如数据列 ID 或名称。

Step4：指定是否只要评估目前选取数据列的影响，或是只要评估数据表中的完整数据集。

Step5：如果选中【当前行】单选按钮，工具便会在对话框中显示结果。当对话框运行时，可以继续测试其他状况。

Step6：如果选中【整个表】单选按钮，工具便会在对话框中显示状态消息，并将两个新的数据行加入原始的数据表中。单击【关闭】按钮即可在工作表中查看完整结果。



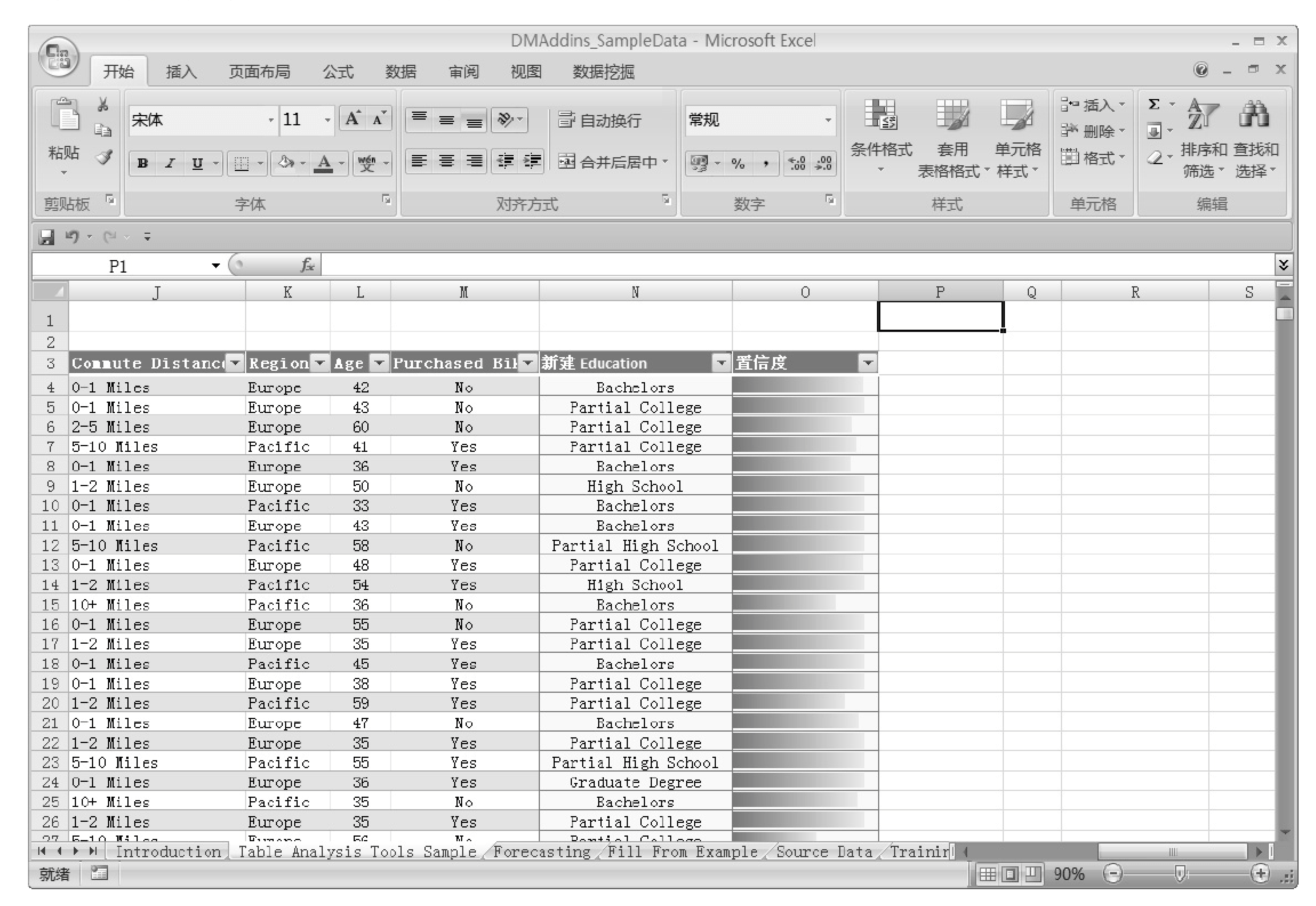
图 22-4 应用场景分析：假设



图 22-5 【假设】对话框

加入数据表中的数据列会包含两种信息类型：变化的预测值及其置信度。置信度表示

预测正确的概率，如图 22-6 所示。



	Commute Distance	Region	Age	Purchased Bik	新建 Education	置信度
4	0-1 Miles	Europe	42	No	Bachelors	
5	0-1 Miles	Europe	43	No	Partial College	
6	2-5 Miles	Europe	60	No	Partial College	
7	5-10 Miles	Pacific	41	Yes	Partial College	
8	0-1 Miles	Europe	36	Yes	Bachelors	
9	1-2 Miles	Europe	50	No	High School	
10	0-1 Miles	Pacific	33	Yes	Bachelors	
11	0-1 Miles	Europe	43	Yes	Bachelors	
12	5-10 Miles	Pacific	58	No	Partial High School	
13	0-1 Miles	Europe	48	Yes	Partial College	
14	1-2 Miles	Pacific	54	Yes	High School	
15	10+ Miles	Pacific	36	No	Bachelors	
16	0-1 Miles	Europe	55	No	Partial College	
17	1-2 Miles	Europe	35	Yes	Partial College	
18	0-1 Miles	Pacific	45	Yes	Bachelors	
19	0-1 Miles	Europe	38	Yes	Partial College	
20	1-2 Miles	Pacific	59	Yes	Partial College	
21	0-1 Miles	Europe	47	No	Bachelors	
22	1-2 Miles	Europe	35	Yes	Partial College	
23	5-10 Miles	Pacific	55	Yes	Partial High School	
24	0-1 Miles	Europe	36	Yes	Graduate Degree	
25	10+ Miles	Pacific	35	No	Bachelors	
26	1-2 Miles	Europe	35	Yes	Partial College	

图 22-6 置信度

也可以在对话框中逐一输入变化值，并以互动的方式查看预测。这与建立单一预测查询（singleton prediction query）相同。预测查询的结果是具有下列信息的输出：预测的成功或失败、预测的值，以及置信度水平。置信度水平会被显示为水平的直方图，柱形越长，表明结果的置信度越高。

第 23 章　Visio 2007 数据透视分析

Visio 2007 的数据透视图表，可以将数据作有结构性的聚类，数据的来源有了更多的选择，除了可依据数据表、数据库，还可以直接连接 cube 来作数据透视图表，如图 23-1 所示。

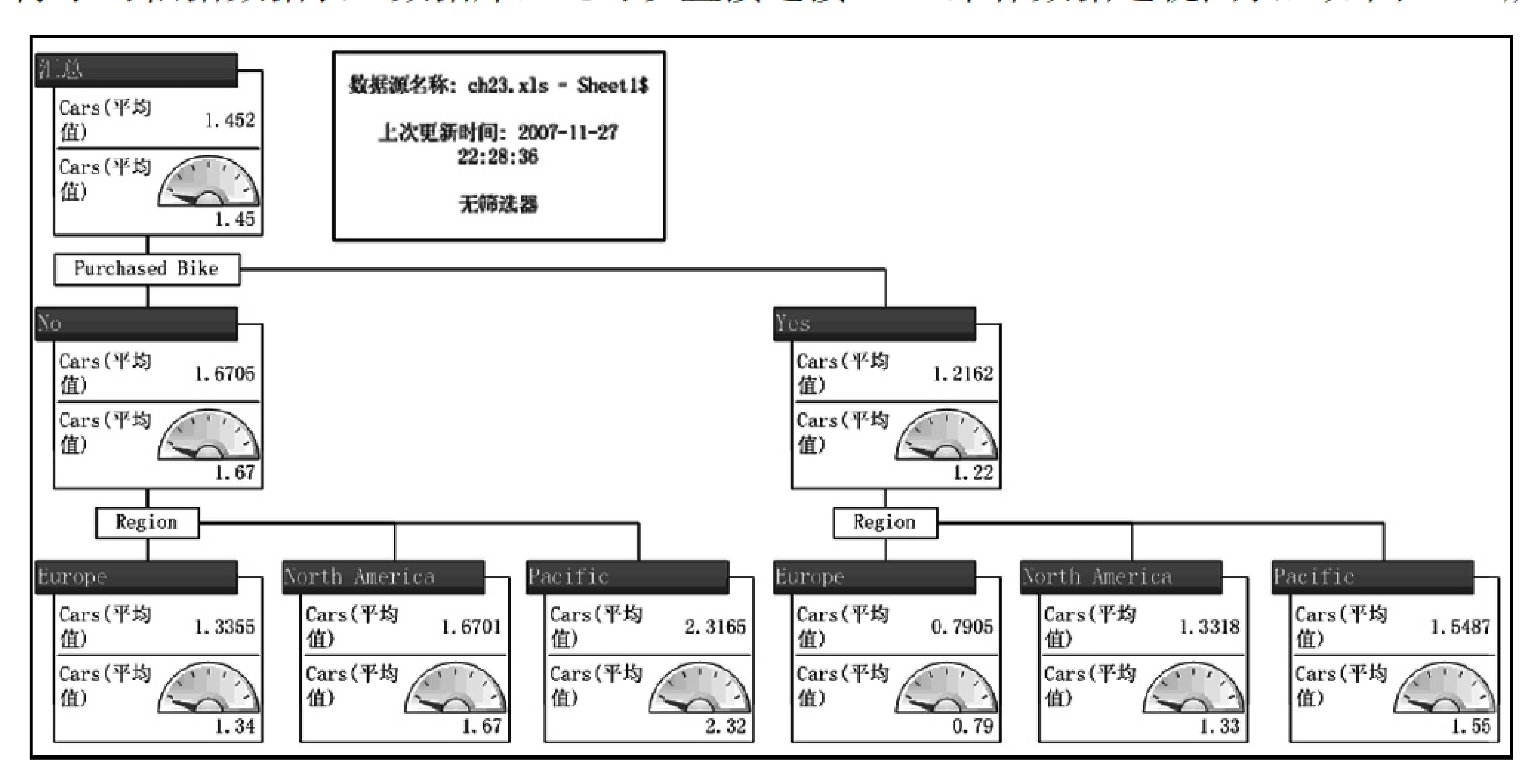


图 23-1　Visio 2007 数据透视图表

Step1：启动 Visio 2007 后，在左边的【模板类别】中选择【商务】，然后在中间的模板图中选择【数据透视图表】，最后在右边视框中单击【创建】按钮。如图 23-2 所示。

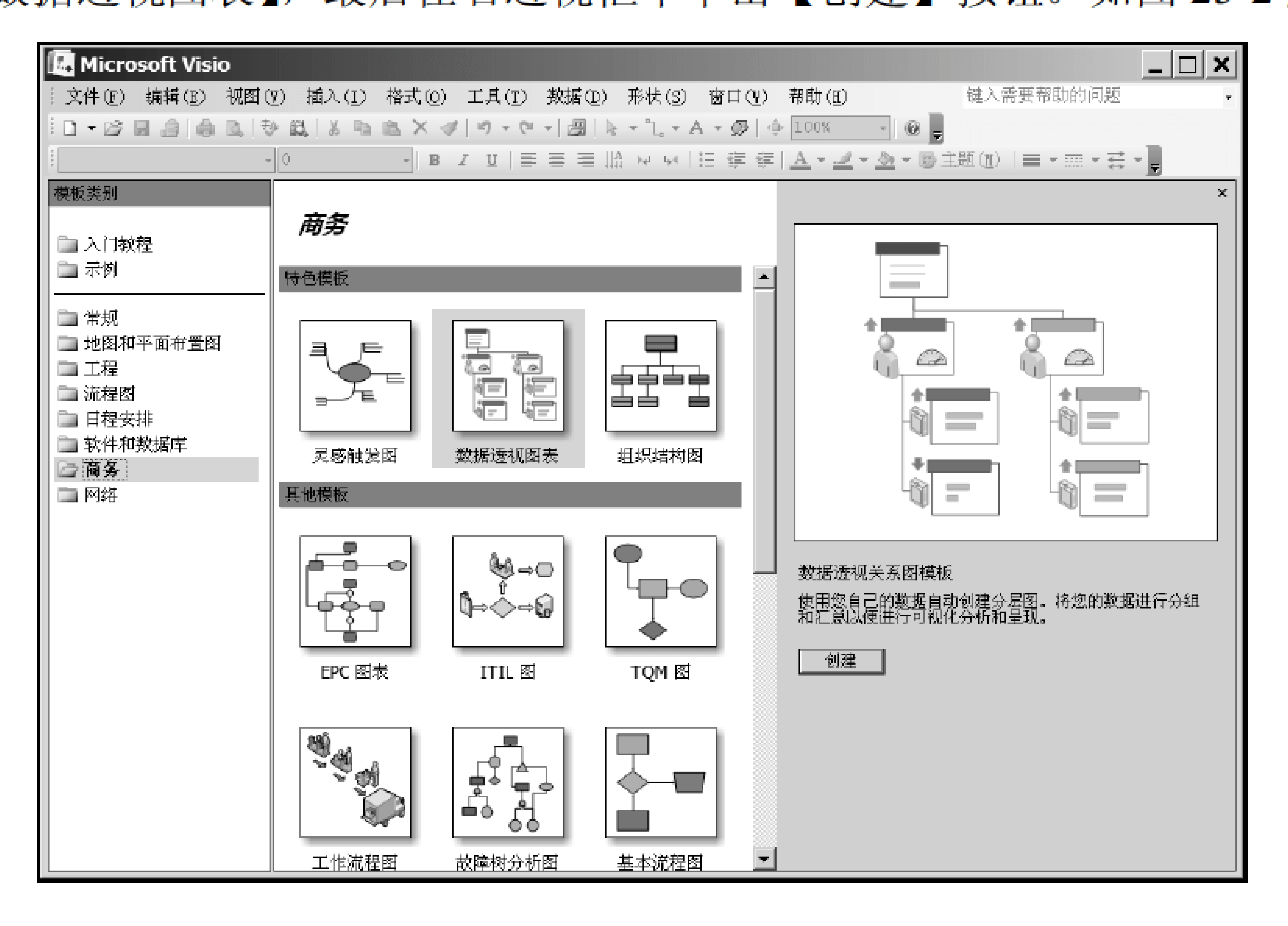


图 23-2　创建数据透视图表

Step2：弹出如图 23-3 所示的【数据选取器】对话框，依据数据来源选择，这里选中【Microsoft Office Excel 工作簿】单选按钮，之后再单击【下一步】按钮。

Step3：在如图 23-4 所示的【要导入的工作簿】下拉列表框中，选择要导入的 Excel 文件。



图 23-3 【数据选取器】对话框



图 23-4 选择要导入的 Excel 文件

Step4：在如图 23-5 所示的【要使用的工作表或区域】下拉列表框中，选择要使用的 Excel 工作表或区域。若数据的第一列包含标题，则选中【首行数据包含有列标题】复选框。如果仅分析一个数据区域，可以单击【选择自定义范围...】按钮进行选择。

Step5：挑选要分析的数据列与数据行，这里选择所有的数据行与数据列。最后单击【完成】按钮，如图 23-6 所示。



图 23-5 选择要使用的 Excel 工作表或区域

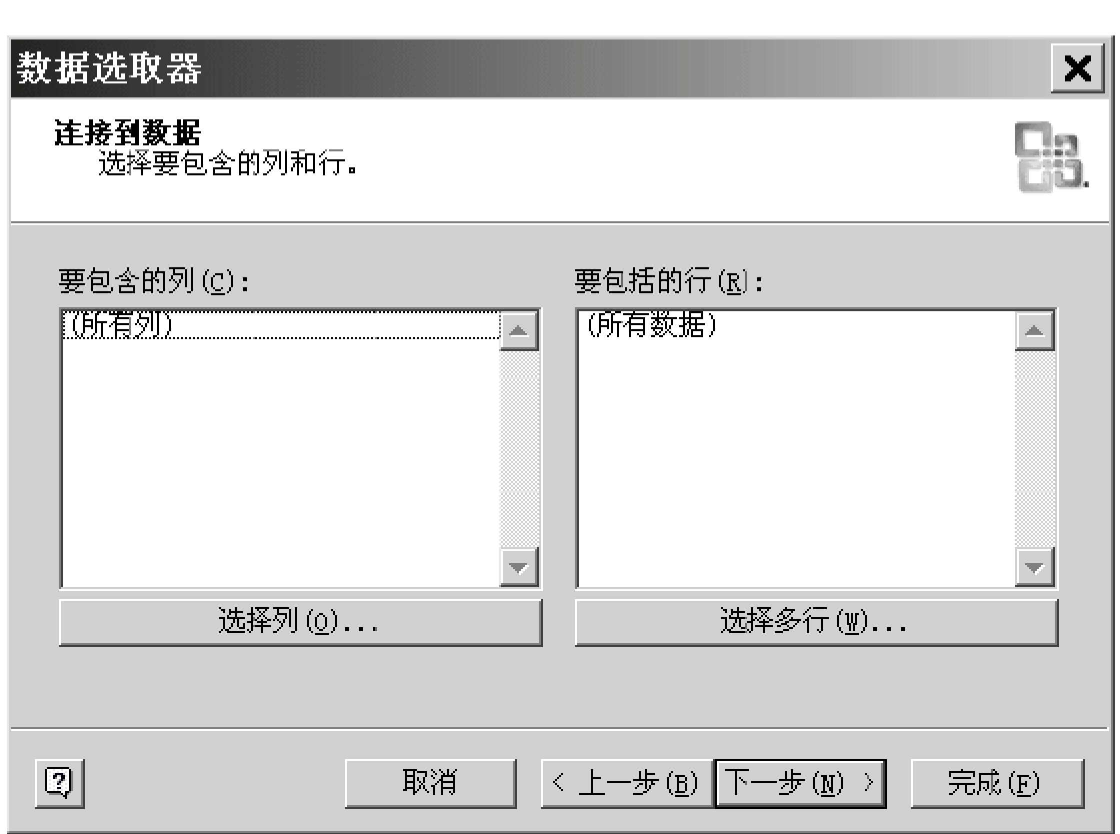


图 23-6 挑选要分析的数据列与数据行

Step6：设定完成要分析的数据来源后，会自动产生一个数据透视图表的基本架构。如图 23-7 所示，1 是数据透视图表的名称，双击可编辑此分析图的名称。2 是数据透视图表的说明，可自行编辑。3 是整个数据透视图表的顶端节点，为整体数据的汇总分析。4 在添加汇总区，系统会自动把数值型字段（即变量）都归入其中。自动预设选中第一个复选框。

Step7：这个步骤主要介绍如何设定要分析的类别与汇总等字段。先取消选中 ID（合计）复选框，因为 ID 字段并不是可以做计算的字段，虽然它是数值型变量；再选中 Cars

（合计）复选框，并将合计改为平均；先选择工作界面的汇总项，再从“添加类别”栏中依次选中要分析的 Purchased Bike 和 Region，如图 23-8 所示。整个数据透视图表的结构就会先按 Purchase Bike 节点分类，再按 Region 分类。

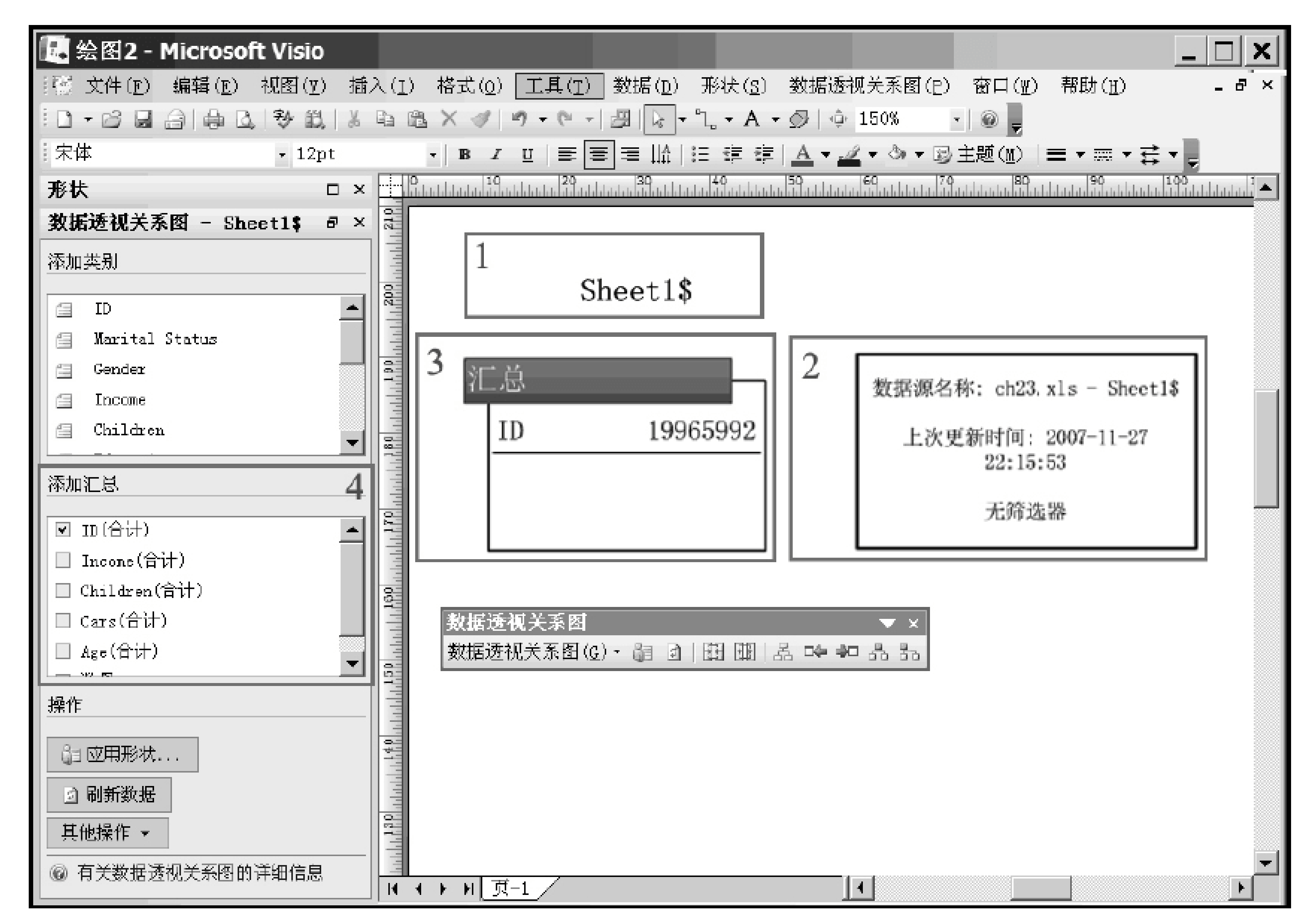


图 23-7　数据透视图表基本架构

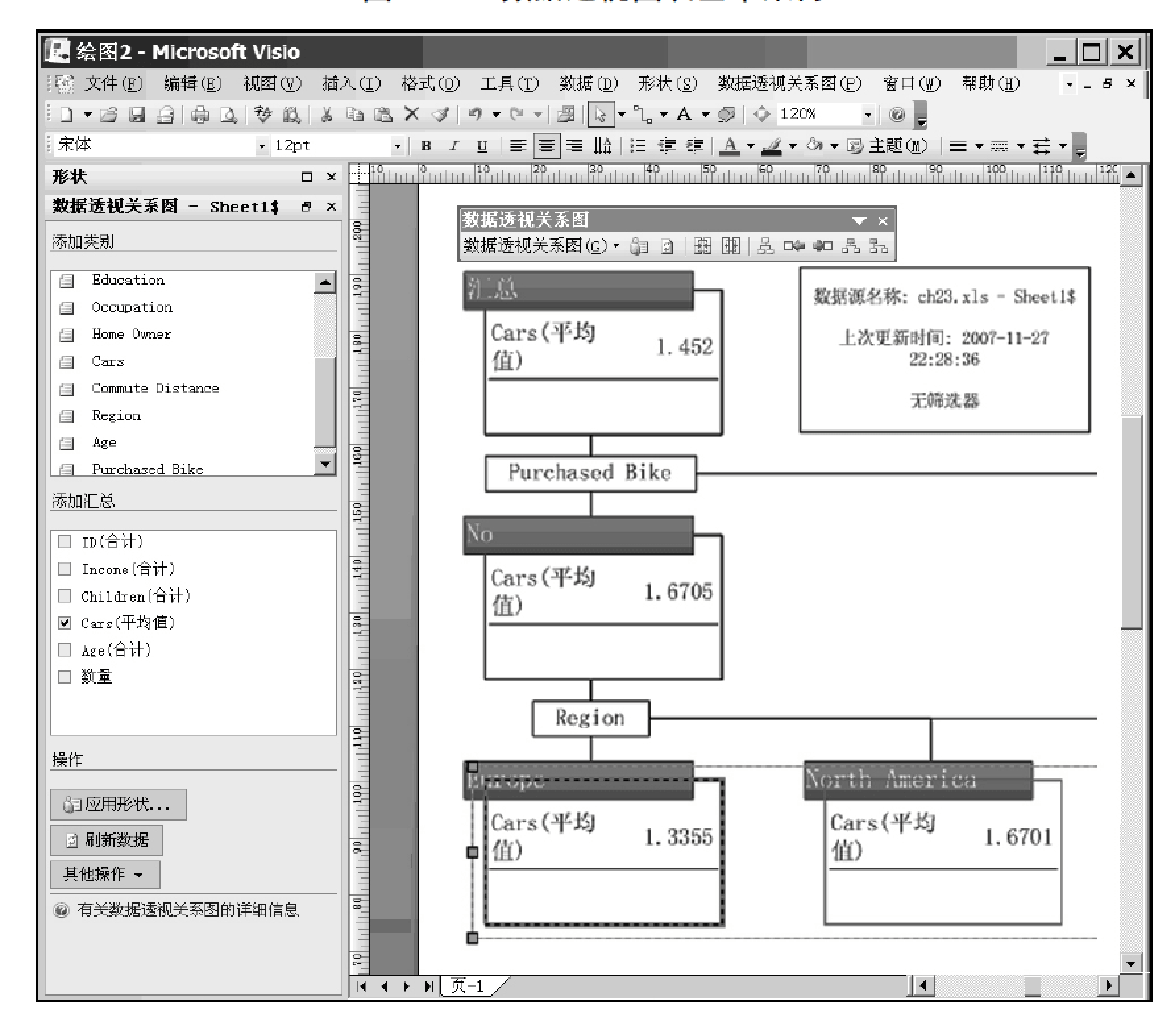


图 23-8　设定要分析的类别与汇总

Step8：接下来介绍如何将数据数值以图标表现。在某一个节点的数值上右击，在弹出的快捷菜单中选择【数据】→【编辑数据图形】命令，如图 23-9 所示。

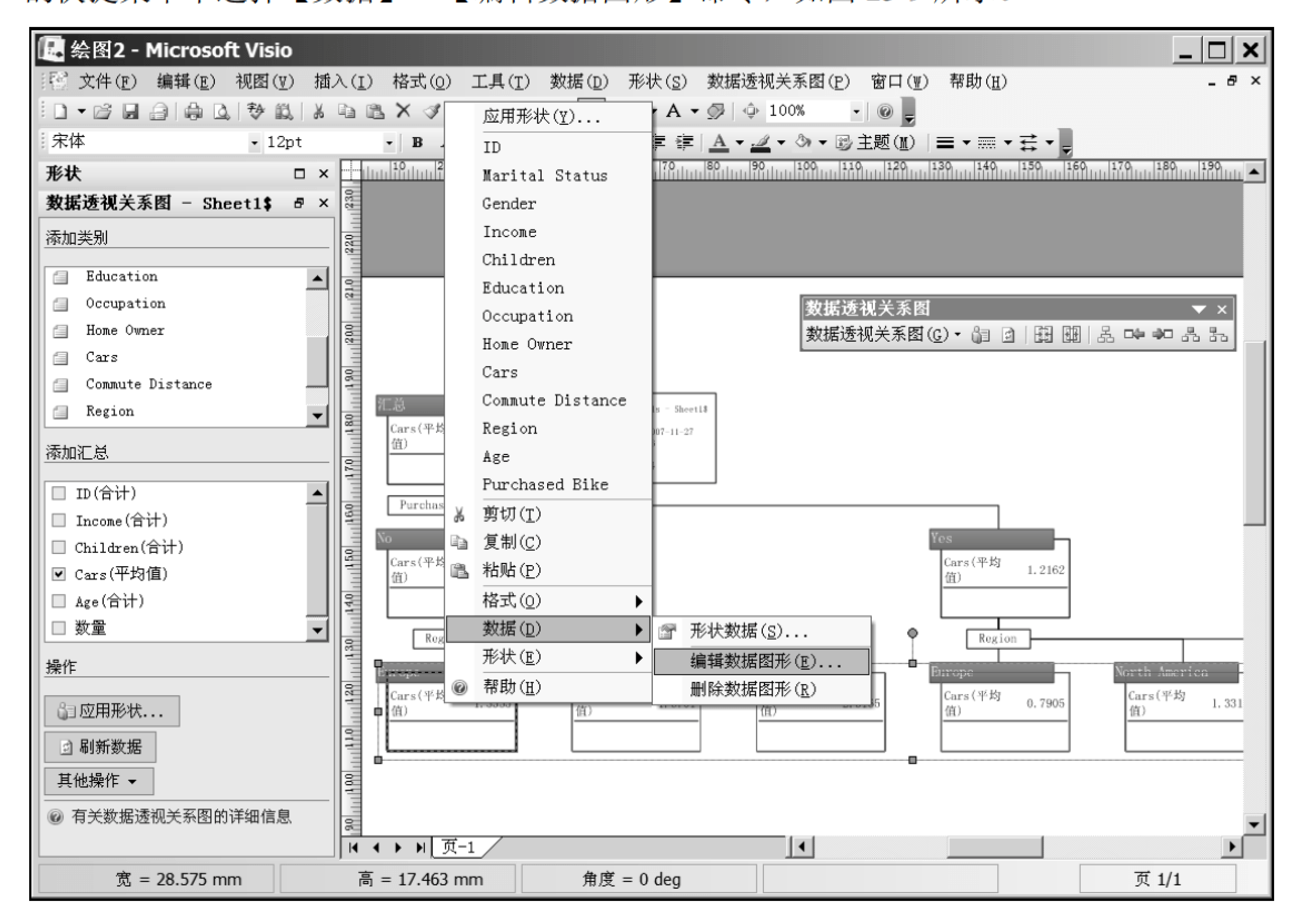


图 23-9 编辑数据图形

Step9：在弹出的如图 23-10 所示的【编辑数据图形】对话框中选择【新建项目】→【数据栏】命令。

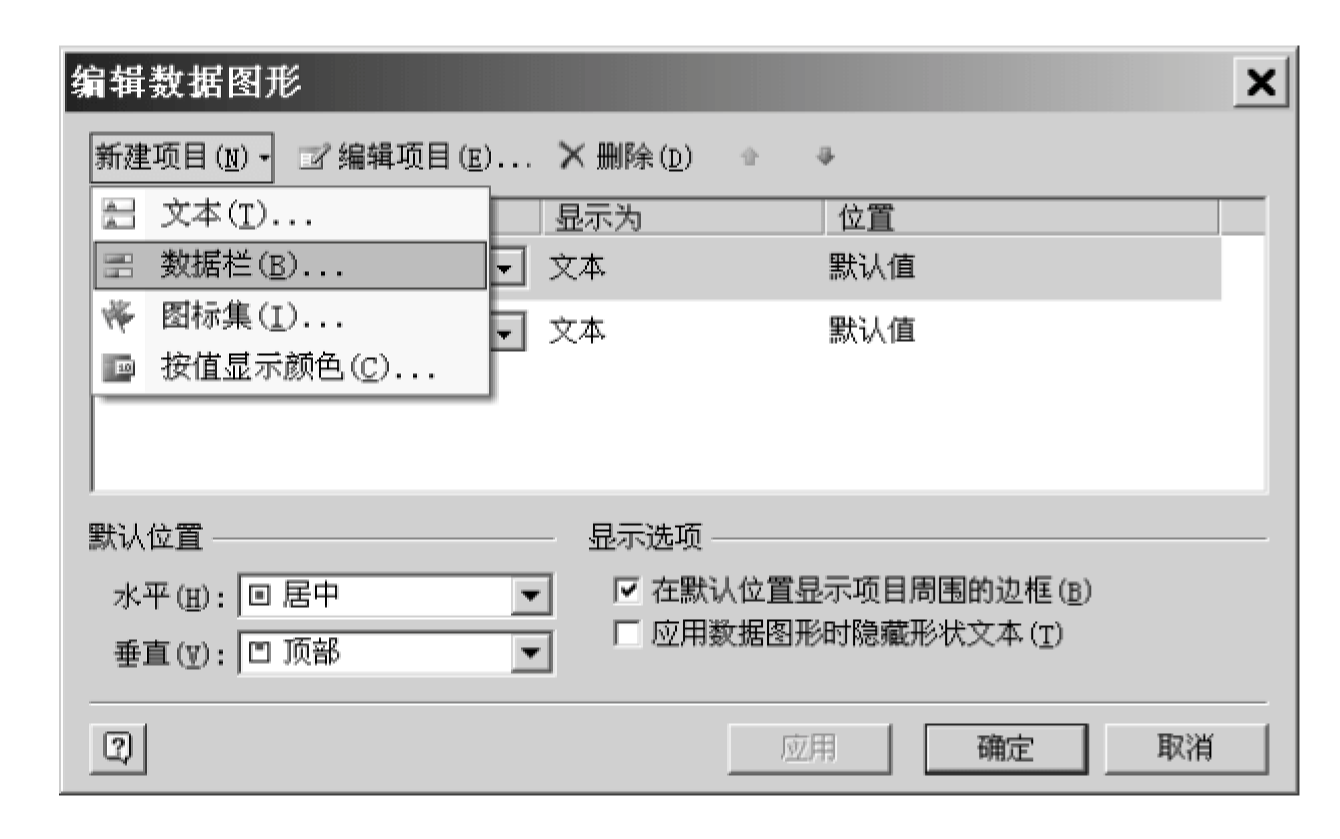


图 23-10 【编辑数据图形】对话框

Step10：在如图 23-11 所示的【新建数据栏】对话框中的【数据字段】下拉列表框中选择 Cars（平均值），在【标注】下拉列表框中选择速度计，最后单击【确定】按钮，即可完

成，如图 23-12 所示。



图 23-11　【新建数据栏】对话框

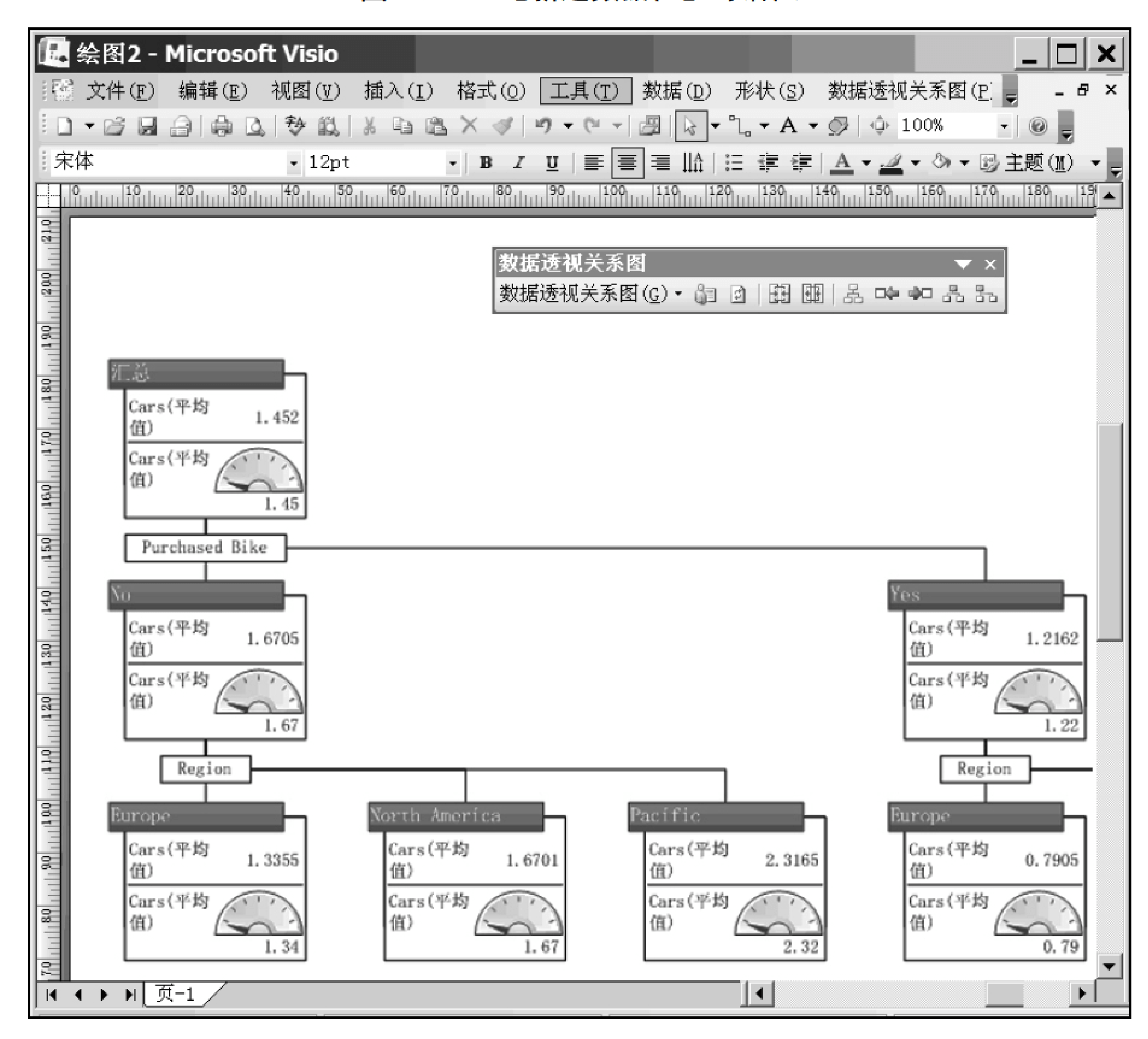


图 23-12　完成后的数据透视图表



第4篇

数据挖掘范例

- 上市公司投资价值分析的挖掘模型
- 信用卡用户信用评测的挖掘模型
- 市场营销与客户细分的挖掘模型

第 24 章　上市公司投资价值分析的挖掘模型

24.1　研究动机与目的

股票作为一种投资工具，其本身代表了股票持有者对公司的所有权，代表取得收益的权利，是对未来收益的支取凭证。股票价格不是它所代表的实际资本价值的货币表现，而是一种资本化的收入。股票价格一般是由股息和利息率两个因素决定的。股票价格=股息/利息率，股票价格与股息成正比例变化，而和利息率成反比例变化。如果某个股份公司的营业情况好，股息增多或预期的股息将要增加，这个股份公司的股票价格就会上涨；反之，则会下跌。

公司在上市一年以后其股价一般取决于其内在价值和成长性因素，质地较好的新上市公司在保证较好业绩和成长性，并伴有高送配题材时，将极大提高公司股价估值水平，在价值投资理念下，股价必然会反映其内在价值。股票走势一般都可以体现业绩增长同股价提升的正相关关系，主营业务发展的好坏，始终是行业和个股股价上涨的原动力；此外股价运行很大程度上也依托于企业的净利润增长空间，利润增长越快，市场中的操作者越能增强其持有信心，即净利润的相对高速增长能调高市场对该股票的预期。每股收益常被用来衡量企业的盈利能力和评估股票投资的风险。如果企业的每股收益较高，则说明企业盈利能力较强，从而投资于该企业股票风险相对也就小一些。因此，投资者处于盈利和避险的考虑往往选择每股收益高的新股进行投资，这样每股收益相对高的新股股价得以市场资金和人气的支持，而展现高于一般的涨幅。可见新股每股收益和其股价运行存在着正相关关系。

24.2　挖掘模型的构建

让计算机通过个股的盈利状况数据来选择股票，可以看作一个数据挖掘中的分类问题。因为知道每支股票过去的盈利状况数据和涨跌走势。可以把这些数据看作一个数据集。如果让机器通过归纳法对这个数据集进行处理，通过训练数据建立模型，进行准确性检验和改进，再应用于新的数据。例如新近公布的上市公司业绩状况，就可以对上市公司进行分类，帮助投资者选择合适的股票，并能够预测将来的股票是否超涨还是处于一般水平。

本范例选取了 2006 年 12 月 31 日 1 110 支中国 A 股的一系列与利润相关的财务指标：主营业务收入、净利润、总资产、股东权益（不含少数股东权益）、每股收益（摊薄净利润）、净资产收益率（净利润）、资产收益率、净利润率和市净率以及这些股票在 2007 年第一季度的涨幅状态（标识个股是否高于平均水平）。

基本的挖掘模型是：先将选取的数据随机分成两个部分，其中训练集占 70%，验证集占 30%。并分别用 Logistic 回归、贝叶斯分类和决策树对其涨幅状态建模。然后用验证集进行检验和比较，如图 24-1 所示。

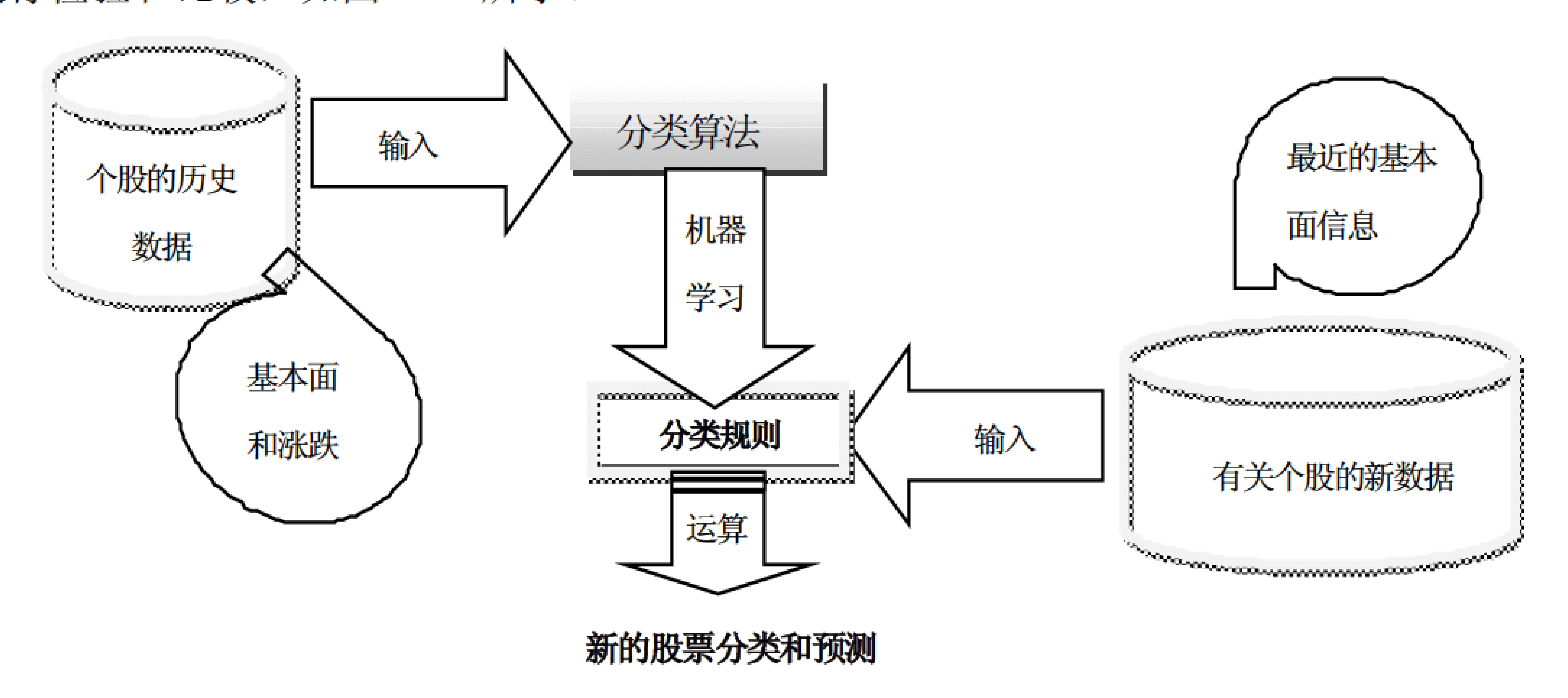


图 24-1　上市公司投资价值分析的挖掘模型框架

24.3　变 量 筛 选

研究目的是发现股票基本面信息与股票价格增长之间的关系，借助公司的基本面信息来选择出股价超额增长的股票，从而在股票市场上获取超额收益。因此指标的选取自然也分成了两个部分：有关公司基本面的指标；有关股票价格增长的指标。

作为对公司股票价格走势进行预测、分类的依据，主要选择了主营业务收入、净利润、总资产、股东权益（不含少数股东权益）、每股收益（摊薄净利润）、净资产收益率（净利润）、净利润增长率、市净率等指标来描述公司基本面信息。各指标的定义有以下几个，选用变量如表 24-1 所示。

主营业务收入：指的是企业（集团）从事某种主要生产、经营活动所取得的营业收入（单位：元）。

净利润：是一个企业经营的最终成果，衡量企业经营效益的主要指标。

总资产：指企业拥有或控制的全部资产。包括流动资产、长期投资、固定资产、无形及递延资产、其他长期资产、递延税项等，即为企业资产负债表的资产总计项。总资产代表了企业的长期偿债能力。

股东权益（不含少数股东权益）：指公司所有者权益合计。不包括公司投资子公司产生的少数股东权益。股东权益衡量的是公司总资产扣除负债之后的余额，是股东在公司清算情况下最终能够获得利益多少的一个指标。

每股收益（摊薄净利润）：计算公式为每股收益 = 除税后净利润/已发行股票数。上市公司如发行可换股债、认股权证、配送新股，则假设其全部行权，在计入可换股债（和/或）认股权证所产生的新股数，计算出新的每股收益称为摊薄后的每股收益。

净资产收益率（净利润）：又称股东权益收益率，是净利润与平均股东权益的百分比。该指标反映股东权益的收益水平，指标值越高，说明投资带来的收益越高。净资产收益率的计算公式是公司税后利润除以净资产得到的百分比率，用以衡量公司运用自有资本的效率。

净利润增长率：即本年净利润减去上年净利润之差再除以上期净利润的比值。净利润是公司经营业绩的最终结果。净利润的增长是公司成长性的基本特征。

市净率：公式为市净率 = 股票市价/每股净资产。净资产的多少由股份公司经营状况决定。

表 24-1　选用变量及其类型说明

指 标 名 称	变 量 类 别	作　　用	注　　释
股票名称	分类变量	不使用	
股票代码	分类变量	样本编号	
主营业务收入	连续变量	输入变量	2006 年第四季度
净利润	连续变量	输入变量	2006 年第四季度
总资产	连续变量	输入变量	2006 年第四季度
股东权益（不含少数股东权益）	连续变量	输入变量	2006 年第四季度
每股收益（摊薄净利润）	连续变量	输入变量	2006 年第四季度
净资产收益率（净利润）	连续变量	输入变量	2006 年第四季度
资产收益率	连续变量	输入变量	2006 年第四季度
净利润增长率	连续变量	输入变量	2006 年第四季度
市净率	连续变量	输入变量	2006 年第四季度
是否为较高增值股票（good）	离散变量	指示变量	较高增值股票为 1，否则为 0

在本研究中，需要根据股票价格的增长情况，将股票分为两类：涨幅高于一般水平的股票和低于一般水平的股票。前者即为具有投资价值的上市公司股票。

由于上证指数采用加权综合价格指数公式计算，主要代表了一些股票的走势，而受成份股公司市值的影响很大。深成指也存在与上证指数十分类似的问题，因此以大盘指数的增长率与个股涨幅相比不妥当。此外，考虑中小投资者的资金有限，选择投资组合时并不是按照拟投资股票的市值进行配置，多采取简单“投资组合”方法，例如将资金等分成若干份，然后分别投入不同的股票。由此，选择 A 股股票的平均涨幅是一种更加合适的做法。

经过计算，在 2007 年第一季度的第一个和最后一个交易日之间，深沪两市有正确资料可查的 1 110 支股票，其平均涨幅为 73.9%。由此以 73.9%为临界值，高于此值的叫做较高增值股票，低于此涨幅的叫较低增值股票。

24.4　决策树模型

利用决策树方法，所有的股票被分成了两层五类。在第一层，利用每股收益摊薄净利润这个指标，分成了三类，而根据 Gini 系数差异最大化分类的原则，将其中两类再分别分为两小类，如图 24-2 所示。

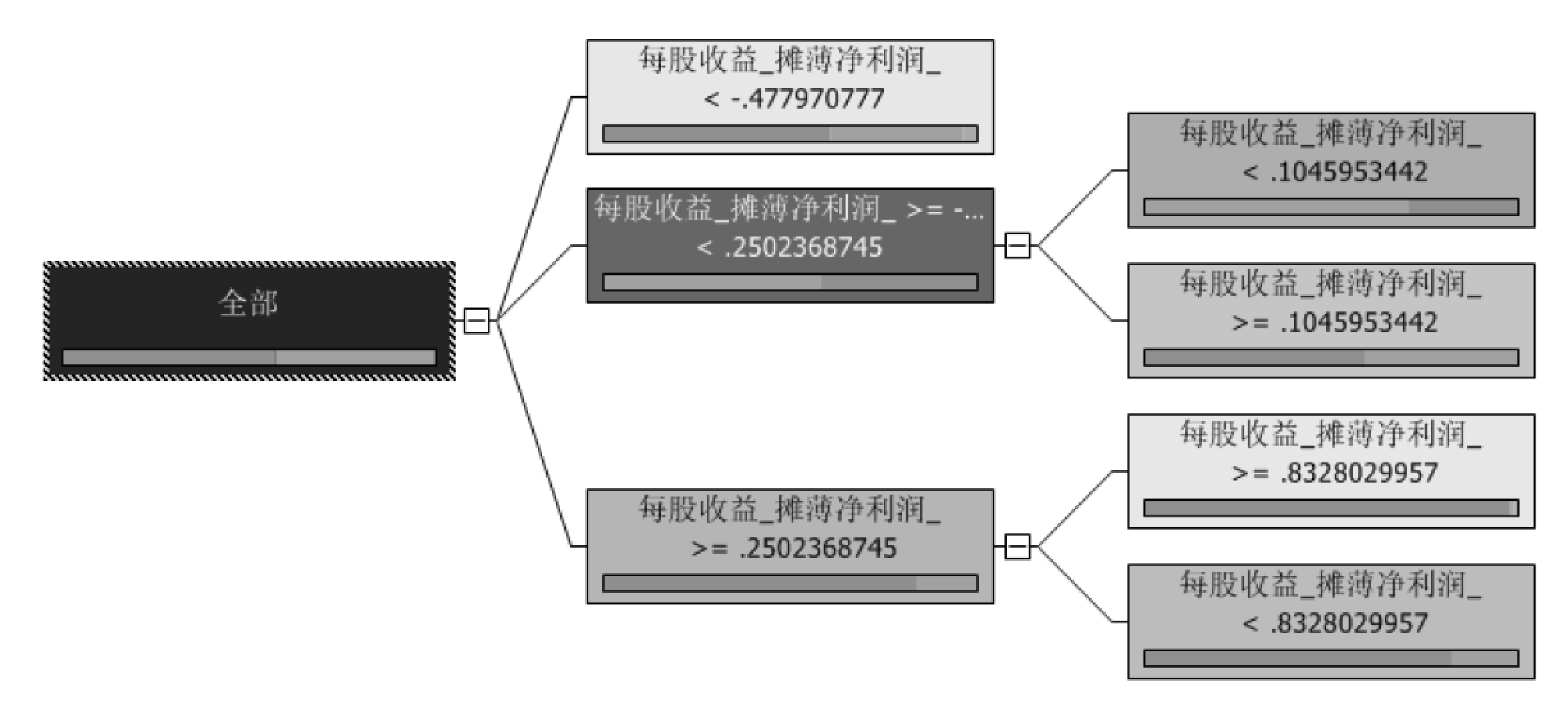


图 24-2　决策树分析结果

现在对每一类的结果进行解释。

第一类是每股收益小于-0.478，这一类属于业绩很差的股票。这一类共有 25 支股票，其中涨幅超过平均值的股票有 9 支，占所有此类股票总数的 36%。而没有涨过平均值的股票有 16 支，占此类股票总数的 64%。

第二类是每股收益小于 0.250 且大于-0.478 的股票，这类股票属于业绩中等的股票，这一类共有 489 支股票，其中涨幅超过平均值的股票有 284 支，占所有此类股票总数的 58%。而没有涨过平均值的股票有 205 支，占此类股票总数的 42%，其中利用 Gini 系数可以把股票进一步划分为两个小类。

第二类的第一小类是每股收益大于-0.478 小于 0.105，这一类属于业绩比较差的股票。这一类共有 283 支股票，其中涨幅超过平均值的股票有 199 支，占所有此类股票总数的 70.31%。而没有涨过平均值的股票有 84 支，占此类股票总数的 29.69%。

第二类的第二小类是每股收益大于 0.105 小于 0.250，这一类属于业绩中等的股票。这一类共有 209 支股票，其中涨过幅度超过平均值的股票有 89 支，占所有此类股票总数的 42.58%。而没有涨过平均值的股票有 120 支，占此类股票总数的 57.42%。

第三类是每股收益大于 0.250 的股票，这类股票属于业绩比较好的股票。这一类共有 263 支股票，其中涨幅超过平均值的股票有 42 支，占所有此类股票总数的 15.97%。而没有涨过平均值的股票有 221 支，占此类股票总数的 84.03%，其中利用 Gini 系数可以把股票进一步划分为两个小类。

第三类的第一小类是每股收益大于 0.250 小于 0.833，这一类属于基本面业绩比较好的股票。这一类共有 237 支股票，其中涨幅超过平均值的股票有 99 支，占所有此类股票总数的 17.73%。而没有涨过平均值的股票有 195 支，占此类股票总数的 82.27%。

第三类的第二小类是每股收益大于 0.833，这一类属于业绩最好的股票。这一类共有 26 支股票，其中涨幅超过平均值的股票有 0 支，占所有此类股票总数的 0%。而没有涨过平均值的股票有 26 支，占此类股票总数的 100%。

从决策树方法可以发现，由于有多重共线性的影响，最终分类的依据标准仅仅从每股收益－摊薄净利润这个指标出发。从结果发现，业绩好的股票在第一个季度的表现并不是太好，第三类的好股票中超过平均收益率的只有很小的比率。而基本面中等，或者偏下的股票反而涨幅比较居前。

24.5　贝叶斯概率模型

贝叶斯方法的结果如表 24-2 所示，在下面说明时仅对表现比较充分的几项进行说明（大于 30%）。

表 24-2　单纯贝叶斯方法各种属性对类标签值的影响

属 性 名 称	属性取值区间	股票超涨的概率/%
股东权益_不含少数股东权益	< 1 860 536 171.9	89
净利润	< 139 785 482.6	92
净资产收益率_净利润_	0.017 5～0.044	31
每股收益_摊薄净利润_	0.007～0.214	70
市净率	< 1.444	31
资产收益率	< 0.007	31

从表 24-2 可以看出，概率越大，说明这个项目越有可能出现超过平均概率的机会。如果股东权益小于 1 860 536 171.9，将会有 89%的概率收益率超过平均值。净利润如果小于 139 785 482.6，将会有 92%的概率，收益率将会超过平均值。如果净资产收益率落在 0.017 5 到 0.044 这个区间上，将会有 31%的概率收益率超过平均值。如果摊薄净利率能够在 0.007 到 0.214 这个区间内，将会有 70%的概率使得收益率超过平均值，如果市净率小于 1.444，则有 31%的概率使得收益率超过平均值，如果资产收益率小于 0.007，也会有 31%的概率使得收益率超过平均值。

利用贝叶斯方法，可以发现最有用的是净利润指标、股东权益指标和每股收益的指标，而且无一例外都表明了在这些指标表现得并不是很好的时候，甚至是十分糟糕的时候，这些股票反而容易上涨，而这些指标如果表现得很好，出现大涨的可能性反而很低。这个发现和决策树的分析结果一致。

24.6　Logistic 回归模型

Logistic 回归的结果如表 24-3 所示，下面仅对表现比较充分的几项进行说明（大于 30%）。

表 24-3　Logistic 回归各种属性对类标签值的影响

属 性 名 称	属性取值区间	非超涨的概率比/%
总资产	11 530 882 881.0～36 789 173 436.8	100
资产收益率	0.079～0.265	65.57
市净率	7.785～23.920	46.48
每股收益_摊薄净利润_	0.419～1.229	88.1
净利润	-1 930 087 093.6～-314 071 851.9	52.41
股东权益_不含少数股东权益_	4 766 887 835.3～15 410 677 353.4	51.33

其中，favor = 1 表示涨幅超过均值，而 favor = 0 表示涨幅没有超过均值。从表 24-3 可以看出，当总资产在 11 530 882 881.0 到 36 789 173 436.8 这一区间内，都没有超过大盘的，从资产收益率这一变量来看，可以发现，资产收益率在-0.214 到-0.028 之间有 65.57%的概率涨幅超过大盘，而资产收益率在 0.079 到 0.265 之间有 65.57%的概率涨幅低于大盘。从主营业务收入这一角度上看，可以发现主营业务收入在 8 894 696 389.1 到 29 005 915 303.0 之间的企业，有 60.71%的概率涨幅会超过大盘。从市净率来看，如果市净率在 7.785 到 23.920 之间的企业，有 46.48%的概率涨幅无法超过大盘，而如果市净率在-14.929 到-1.575 之间的企业，有 41.41%的概率涨幅可以超过大盘。从每股收益_摊薄净利润这个角度上看，区间在-0.860 到-0.050 的企业会有 88.1%的概率涨幅会超过大盘。而摊薄净利润区间在 0.419 到 1.229 之间的企业，会有 88.1%的概率涨幅会低于大盘。从净利润来看，如果净利润区间在 623 344 541.5 到 2 239 359 783.2 的企业，有 52.41%涨幅会超过大盘，而净利润区间在-1 930 087 093.6 到-314 071 851.9 的企业，有 52.41%的概率涨幅会劣于大盘。从股东权益上看，如果股东权益在 4 766 887 835.3 到 15 410 677 353.4 区间上，有 51.33%的概率优于大盘。

利用 Logistic 回归分析的结果和决策树与贝叶斯分类类似，除了净利润这个指标表现相反之外，大部分业绩指标都是好的时候倾向于涨幅不大，而差的时候涨幅居前。

24.7　预测准确度比较

利用测试集，SQL Server 2005 向模型中逐渐加入样本信息，同时产生对该样本类标签值是否能够获得超额收益的预测。理论上说，随着样本量的加大（即在精确图表中，横轴的坐标值不断上升），预测的准确度会有所提高。如图中绿色的线条（上侧折线）代表理想模型的预测准确度轨迹。而随机猜测的精度为蓝色线条（下侧直线）。而代表模型的预测准确度轨迹的红线越贴近绿线（即靠近上侧折线），预测效果越好。

对红色轨迹线积分后的面积除以对蓝色的随机模型轨迹线积分后的面积，就是该模型的增益信息量，增益信息量越大越好。从图 24-3～图 24-5 可以看出，三个模型和理想模型的差距都较大，增益信息最多的模型是单纯贝叶斯模型，其次是 Logistic 回归模型，决策树模型所含增益信息明显少于前两者。

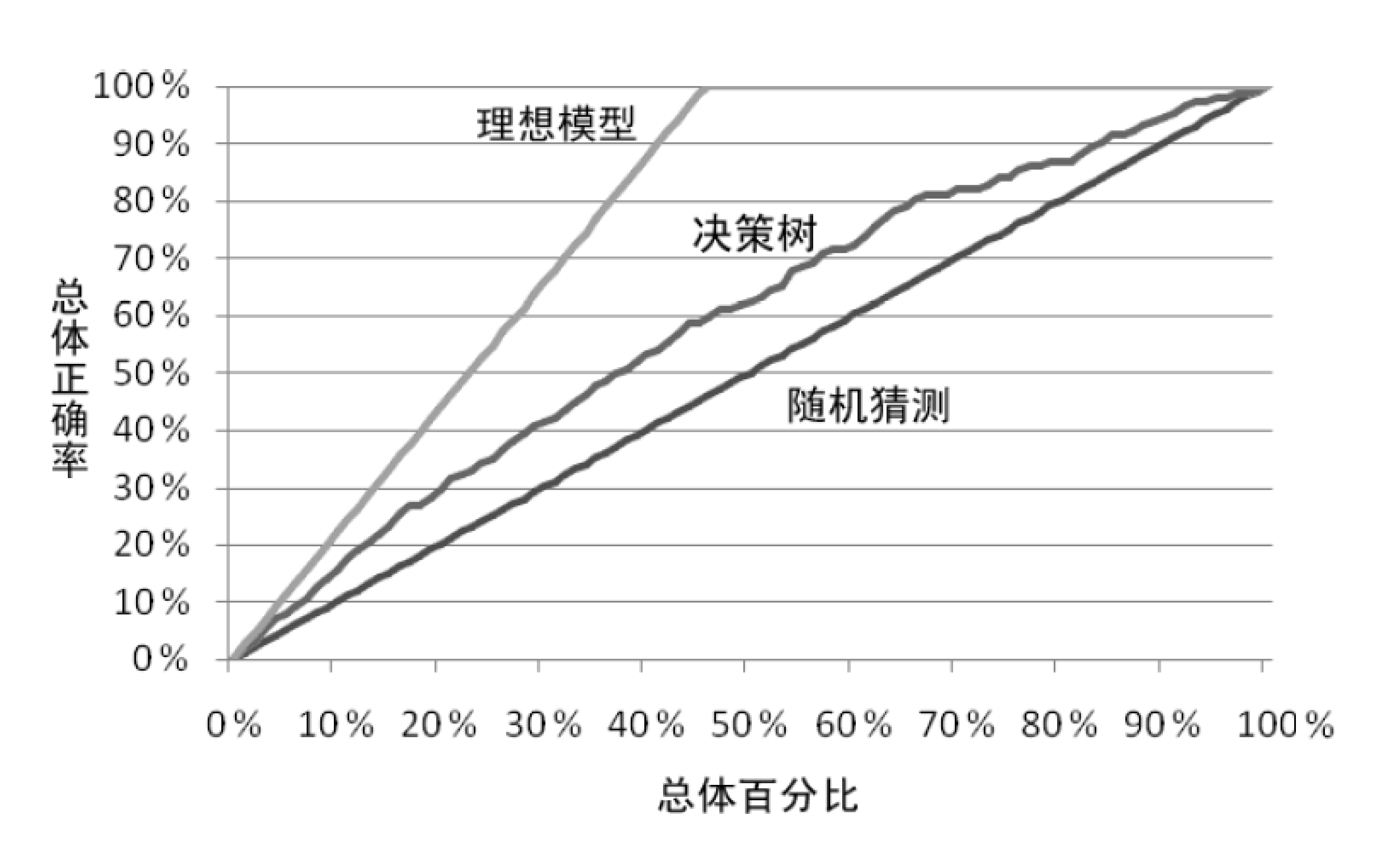


图 24-3　决策树精确图表，增益信息 117.89 %

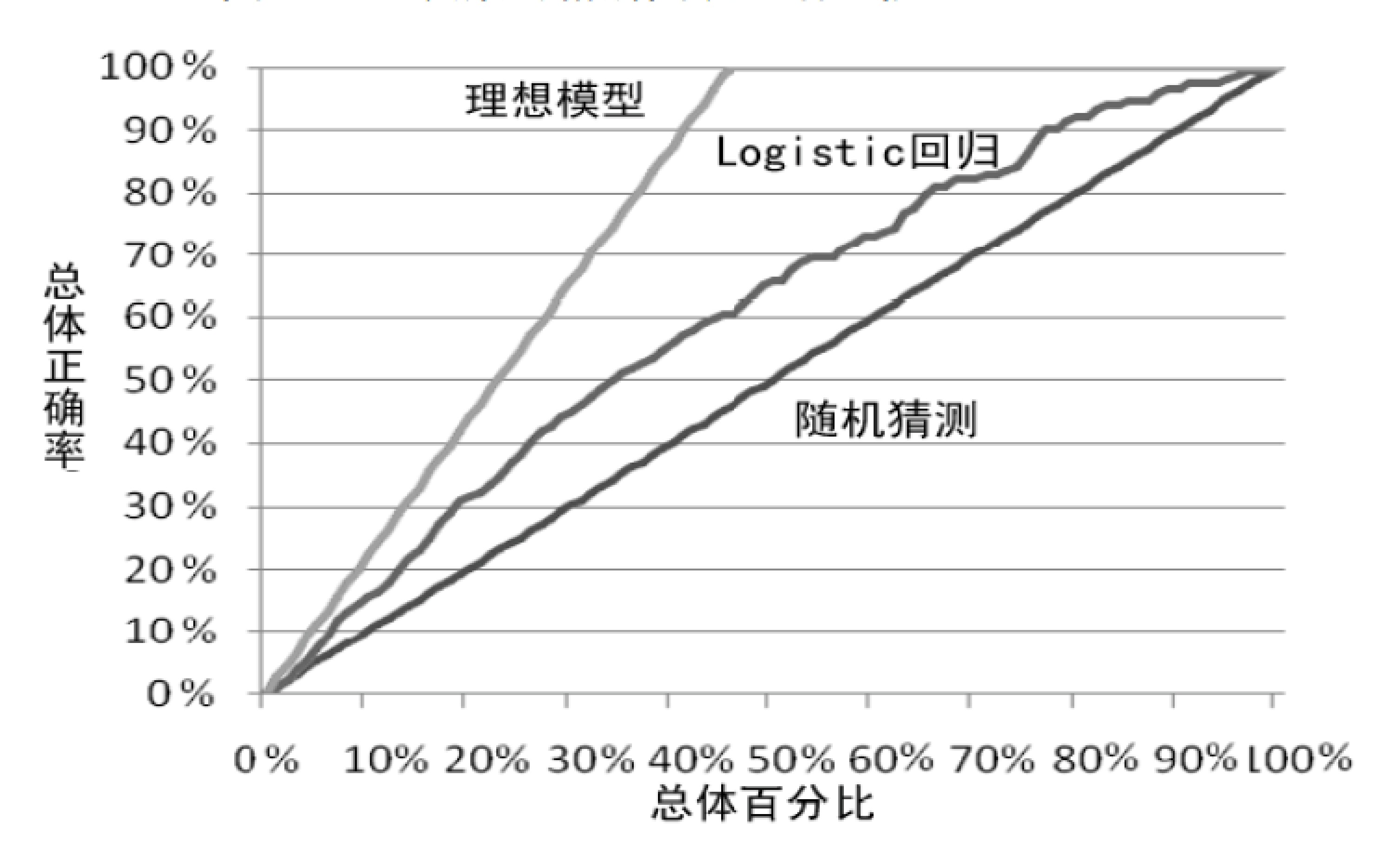


图 24-4　Logistic 回归精确图表，增益信息 121.10 %

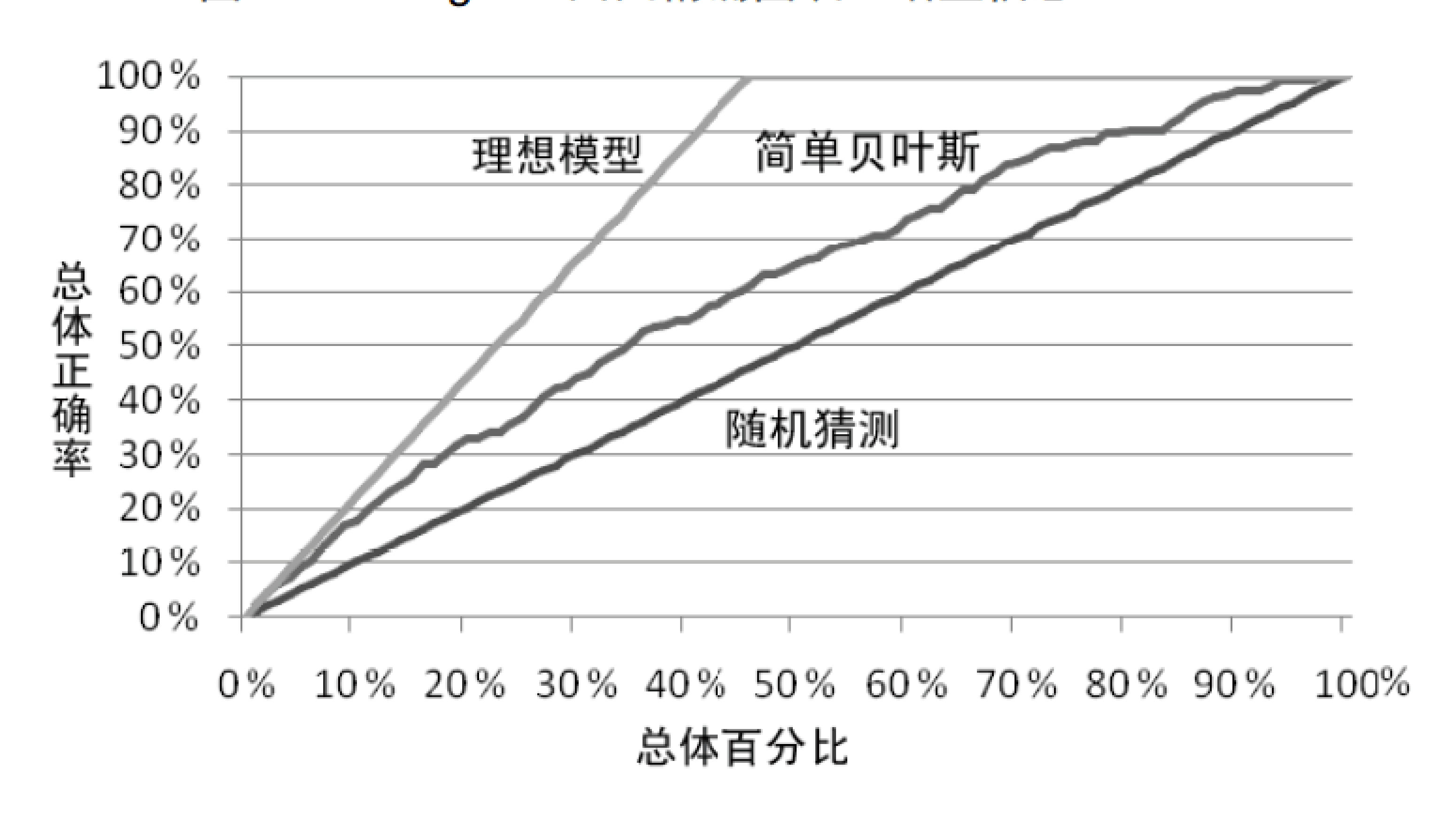


图 24-5　简单贝叶斯分类精确图表，增益信息 121.47%

利用测试集数据，还可以进行预测成本比较，即建立分类矩阵，体现出模型预测某支股票会超涨或不会超涨时的准确率。三种方法的分类矩阵如表 24-4～表 24-7 所示。

表 24-4　决策树方法的分类矩阵

决　策　树	比率/%	个　　数
总体正确率	63.25	210
总体误判率	36.75	122
分类矩阵	0（Actual）	1（Actual）
0	60.00	32.89 %
1	40.00	67.11 %

表 24-5　Logistic 回归的分类矩阵

Logistic 回归	比率/%	个　　数
总体正确率	65.66	218
总体误判率	34.34	114
分类矩阵	0（Actual）	1（Actual）
0	71.67	52.63 %
1	28.33	47.37 %

表 24-6　简单贝叶斯方法的分类矩阵

简单贝叶斯分类	比率/%	个　　数
总体正确率	63.25	210
总体误判率	36.75	122
分类矩阵	0（Actual）	1（Actual）
0	60.00	32.89 %
1	40.00	67.11 %

表 24-7　三种方法优劣比较

方　　法	总体正确率	预测投资价值股正确率	预测无投资价值股正确率
Logistic 回归	最优	最优	最差
简单贝叶斯分类	次优	次优	最优
决策树	最差	次优	最优

由表 24-7 可知，三种方法中，Logistic 回归有着微弱的优势。但是，选股者完全不借助其他信息，直接购买逻辑模型分类得出的投资价值股时，正确率只有 47.37%，而此时用简单贝叶斯方法和决策树方法进行选择，准确率将达到 67.11%。但 Logistic 回归模型并非一无是处，很多时候，选股者事先已综合了各种信息罗列出了一个初步的选股清单，这种情况下，当此算法判定某股票具有投资价值，有 71.67%的正确率，显著高于其他两种方法（60.00%）。

简单贝叶斯分类能更清晰地反映出各种指标对于个股涨幅的影响，而且，简单贝叶斯分类的增益信息含量显著高于决策树模型。所以，在实际操作时，应该将简单贝叶斯分类和 Logistic 回归两种方法混合使用。用前者选出投资价值股，用后者选出事前股票列表中的无投资价值股。

第 25 章　信用卡用户信用评测的挖掘模型

25.1　研 究 背 景

中国台湾地区信用卡发卡量持续猛增，截至 2005 年底，岛内信用卡持卡总人数近 880 万人，人均持有 4 张信用卡。信用卡泛滥导致呆账缺口惊人，2003 年我国台湾地区信用卡发卡机构呆账转销金额达 263 亿元新台币（单位下同），其中只收回 41.5 亿元，实际损失高达 221.5 亿元。日趋严重的信用卡欠债问题给社会带来了巨大冲击。受此影响，2006 年上半年岛内信用卡消费出现大幅衰减。据台“行政院主计处”最新公布的统计数据显示，2006 年 1 月至 6 月岛内民众使用信用卡签账金额为 6 961 亿元，比去年同期减少 118 亿元；预借现金金额为 520 亿元，比去年同期大减 568 亿元，显示“卡债”风暴已经严重影响了信用卡消费。

近年来，随着各种信用卡的泛滥，有报道称台湾地区一位“天王级持卡人”一人拥有 160 张信用卡的传奇，他们有些背负巨额信用卡债务，甚至连最低应缴还金额都付不出的“卡奴”也越来越多。根据银行业估计，目前在台湾地区 1 100 万经济人口（有收入或有收入能力者）中，约有 70 万“卡奴”，即每百名经济人口中约有 6 个人是“卡奴”。

作为发行信用卡的银行，如何使自己发行的信用卡盈利最大化，留住忠诚客户，针对不同客户推出差异化服务，减少呆账坏账，都是银行对信用卡客户进行风险管理和营销的一部分。

信用卡主要有以下风险：

信用风险（credit risk），因持卡人信用不良而产生的拒付风险。表现为持卡人由于经济情况恶化，无力还款，使银行贷款无法收回，形成呆账损失的可能性，从而引发信用风险。

欺诈风险（fraud risk），因诈骗所产生的风险，交易非为持卡人所授意或使用。信用卡及卡上信息被盗取后使用，一般来说，损失由发卡银行承担。

作业风险（operation risk），因管理和作业流程上的操作不当而产生的风险。在实际工作中，有的工作人员有章不循、违规操作，从而造成不应有的风险。

客户信用风险评估及动态调整：客户初始信用风险评估是当客户申请新的信用卡时，通过客户填写的基本信息，系统给出的一个建议性的初始信用等级。以客户的历史数据为输入，表现为客户的初始信用等级分布特征，采用数据挖掘技术建立模型，来预测新客户的初始信用等级。客户的信用等级是一个动态变化的过程，应该随着客户消费行为作相应的调整。根据客户的特征数据（ 客户基本信息）、客户的交易数据以及一些有意义的汇总数据，构建动态的信用风险评估模型，让银行了解客户当前信用等级的分布特征，并预测下一阶段（月）客户的信用风险趋势。

25.2 研究动机

基于上述研究背景，本研究得出以下研究动机：

由于信用卡市场蓬勃发展，在信用卡如此普及的情况下，衍生出的商机成为各大银行的焦点，但是同时也潜在着一定的风险。使用信用卡的顾客有数百万笔的庞大数据量，而且每个持卡人的信息（如收入、家庭、居住地点等）对银行进行风险控制很重要。面对如此庞大的数据，需用数据挖掘的技术配合相关的统计方法去分析数据并从中挖掘信用卡市场中的顾客群，对不同的用户提供不同的策略，指引信用卡向着健康的方向发展。

25.3 研究目的

由于如今计算机运算能力的跃进，以及数据储存技术的进步，使得数据挖掘（data mining）成为近年来数据库应用领域中相当热门的议题。数据挖掘技术近年来逐渐受到重视，不仅是因为企业或研究社会的机构单位将它用在描述数据型态与结构上，更重要的是，此技术将管理技巧与管理行为引入，许多文献中都有应用数据挖掘技术成功的例子，其中包括金融业、电信业、零售商、直销营销、制造业、医疗保健及制药业等。

本研究利用数据挖掘技术来为信用卡业带来更深入的信息，找出不同类型的用户，以提供在决策上的判断依据，故利用统计抽样方法结合数据挖掘技术，以聚类分析、决策树分析等统计相关分析方法，针对样本数据进行相关分析，并建立模型，将现有顾客数据加以分群，找出各群中不同特性之分布情形，借以从大量的顾客数据中发掘出信用卡市场的优质客户群，以提供相关信息给业者，并协助业者开发各种产品，以满足各式各样的顾客，以此提升市场占有率。

25.4 Excel 2007 构建数据挖掘模型

25.4.1 决策树分析

在做数据挖掘模型之前，先生成一个新序列，将每笔资料从 1 开始，生成以等差为 1 递增的递增数列，命名为序列，以“序列”作为数据挖掘模型的索引键。在发行信用卡，首先对客户进行初步判断，看其是否具有瑕疵，对瑕疵客户要谨慎，所以首先以“瑕疵户”为因变量，建立不同变量的决策树模型。

1．模型的建立

模型 DT-1 的自变量选取 SQL 给出的建议变量作为自变量，例如呆账、借款余额、拒往记录、年龄模拟、强制停卡记录、血型、职业。而且对是否为瑕疵户的影响的大小分别

为：强制停卡记录>职业>血型>年龄模拟>借款余额>呆账>拒往记录，考虑到这样做有可能会产生过度拟合，决策树过于庞大，故模型 DT-2 自变量选取对瑕疵户影响最强的前四个变量作为自变量，如图 25-1 所示。



图 25-1　SQL 的建议变量

2．模型的决策树分析

DT-1 的树型如图 25-2 所示：有强制停卡记录的用户都是瑕疵户；由于模型比较庞大，选择几条树枝作为解释，没有强制停卡记录，职业为 12（销售职），血型为 2（B 型）的也为瑕疵户。没有强制停卡记录，职业为 12（销售职），呆账为 2（没有呆账记录的人）的也为瑕疵户。

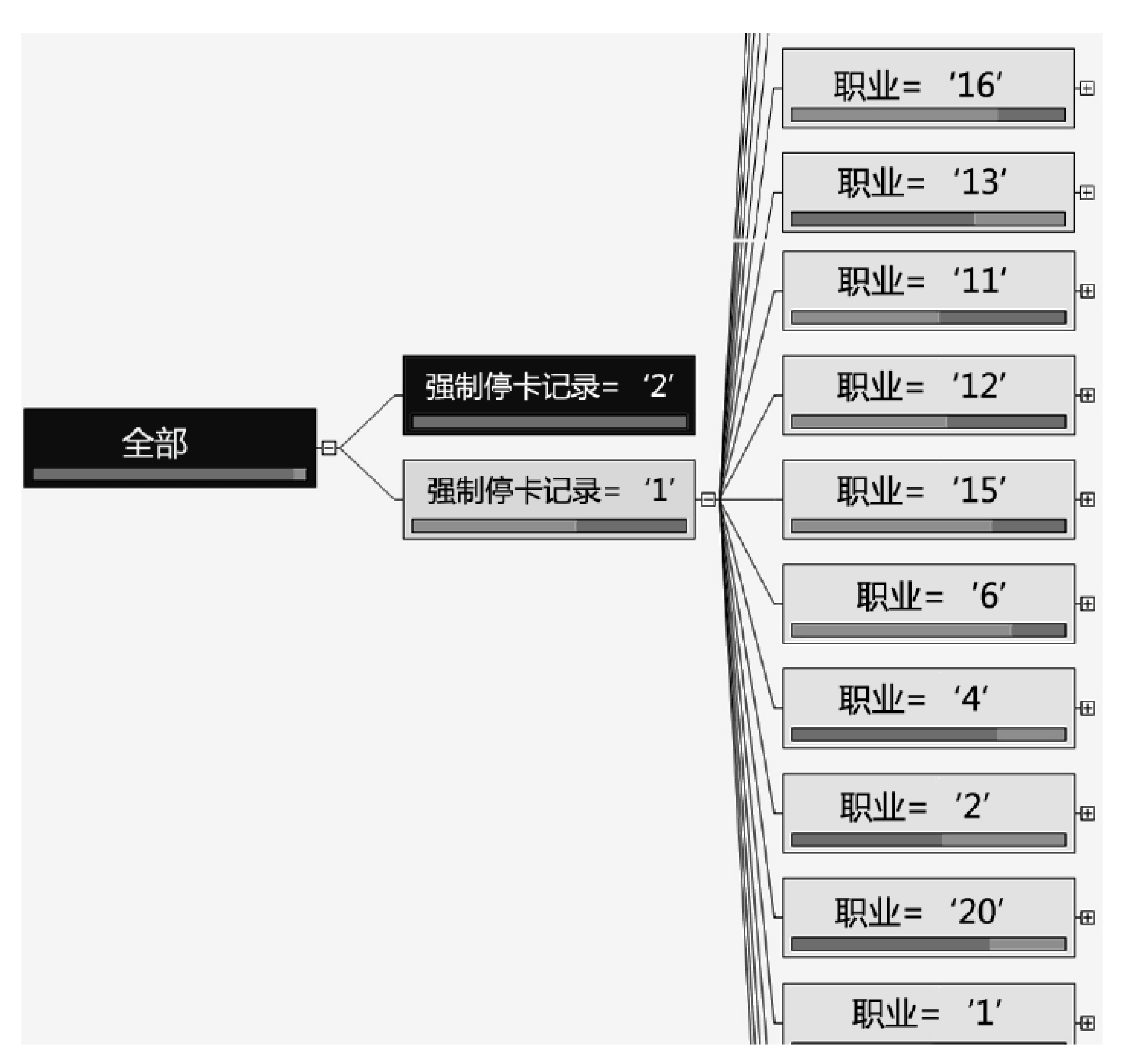


图 25-2　DT-1 树型

DT-2 的树型如图 25-3 所示：有强制停卡记录的用户都是瑕疵户；由于模型比较庞大，选择几条树枝作为解释，没有强制停卡记录，职业为 12（销售职），血型为 2（B 型）的也为瑕疵户。没有强制停卡记录，职业为 11（事物职），年龄模拟为 36，血型为 2（B 型）的也为瑕疵户。

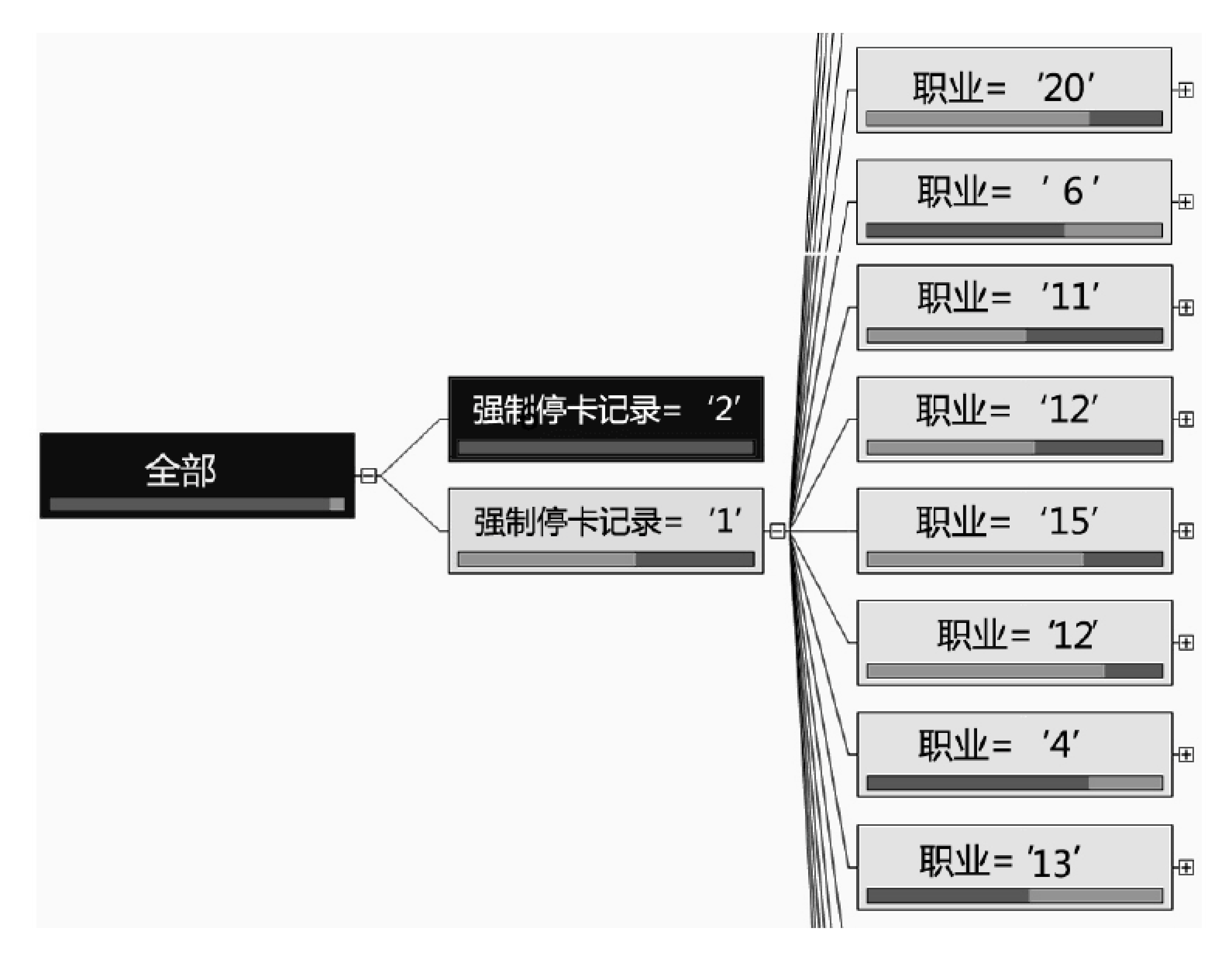


图 25-3 DT-2 树型

3．精确度比较分析

（1）提升图分析

从图 25-4 可以看出，两个模型的拟合精度都很好，在总体百分比为 50%的时候，两个模型的预测率都是 100%。

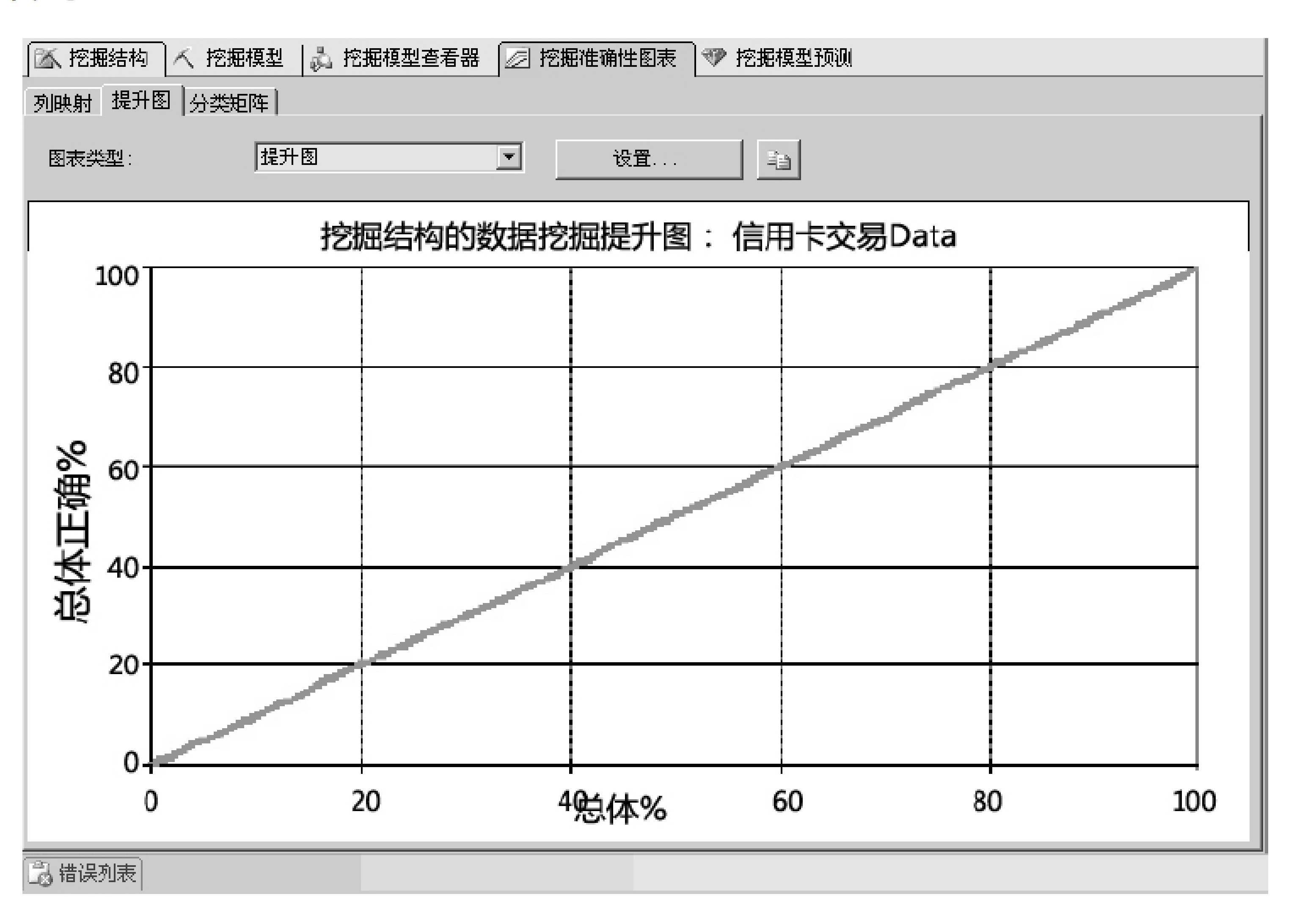
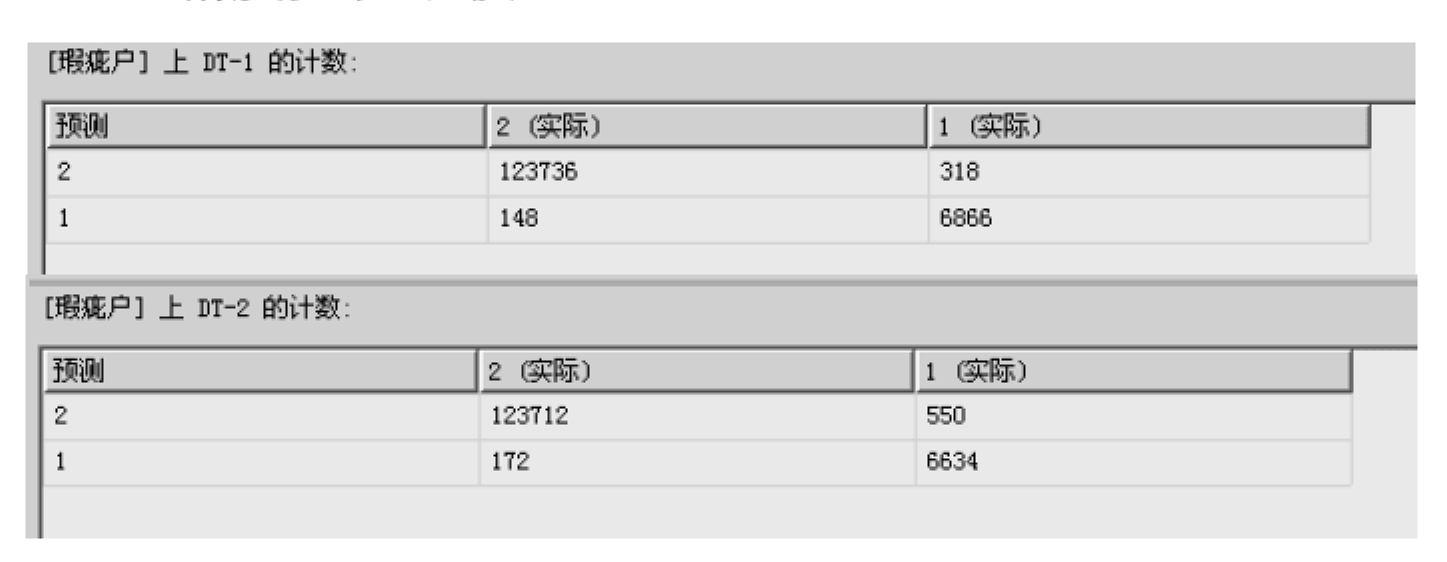


图 25-4 提升图

（2）分类矩阵分析

从图 25-5 也可以看出，瑕疵户取 2 时，模型 DT-1 的正确预测数 123 736 大于 DT-2 的正确预测数 123 712；瑕疵户取 1 时，模型 DT-1 的正确预测数 6 866 大于 DT-2 的正确预测数 6 634，故模型 DT-1 精度要优于模型 DT-2。



[瑕疵户] 上 DT-1 的计数:

预测	2（实际）	1（实际）
2	123736	318
1	148	6866

[瑕疵户] 上 DT-2 的计数:

预测	2（实际）	1（实际）
2	123712	550
1	172	6634

图 25-5　分类矩阵

4．预测分析

报告截取了一小部分数据的事后预测图，分别用了两个模型进行预测，可以看出两个模型大部分的预测是相同的，如图 25-6 所示。



dt1	dt1-p
2	2
2	2
2	2
2	2
2	2
2	2
2	2
2	2
2	2
2	2
2	2
2	2
1	1
2	2
2	2
2	2
2	2
2	2

图 25-6　事后预测图

25.4.2　聚类分析

聚类分析就是通过分析样本数据库中的数据，为每个类别做出准确的描述，或建立分类模型，或挖掘出分类规则，然后用这个分类规则对其他记录进行分类。分类模型也可用于预测，根据已经分好类的资料来研究它们的特征，然后根据这些特征对其他未经分类的或新的数据做预测。例如，将信用卡申请者的风险属性区分为高度风险申请者、中度风险申请者及低度风险申请者。根据经验和数据本身的特征将信用卡持卡人分别分为四类。

1. 分类关系图

图 25-7 所示为分类关系图。

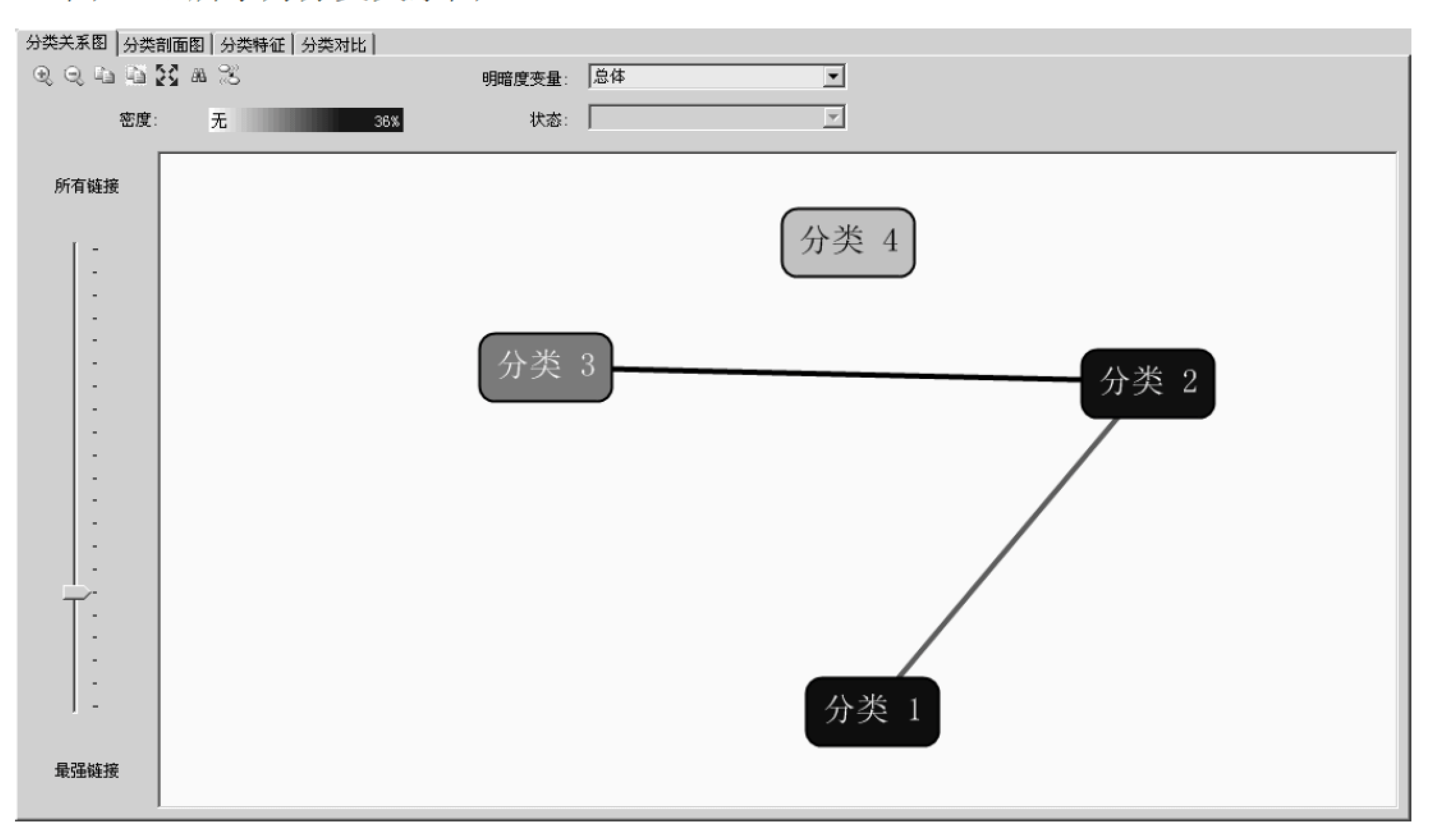


图 25-7　分类关系图

2. 分类剖面图

图 25-8 所示为分类剖面图。

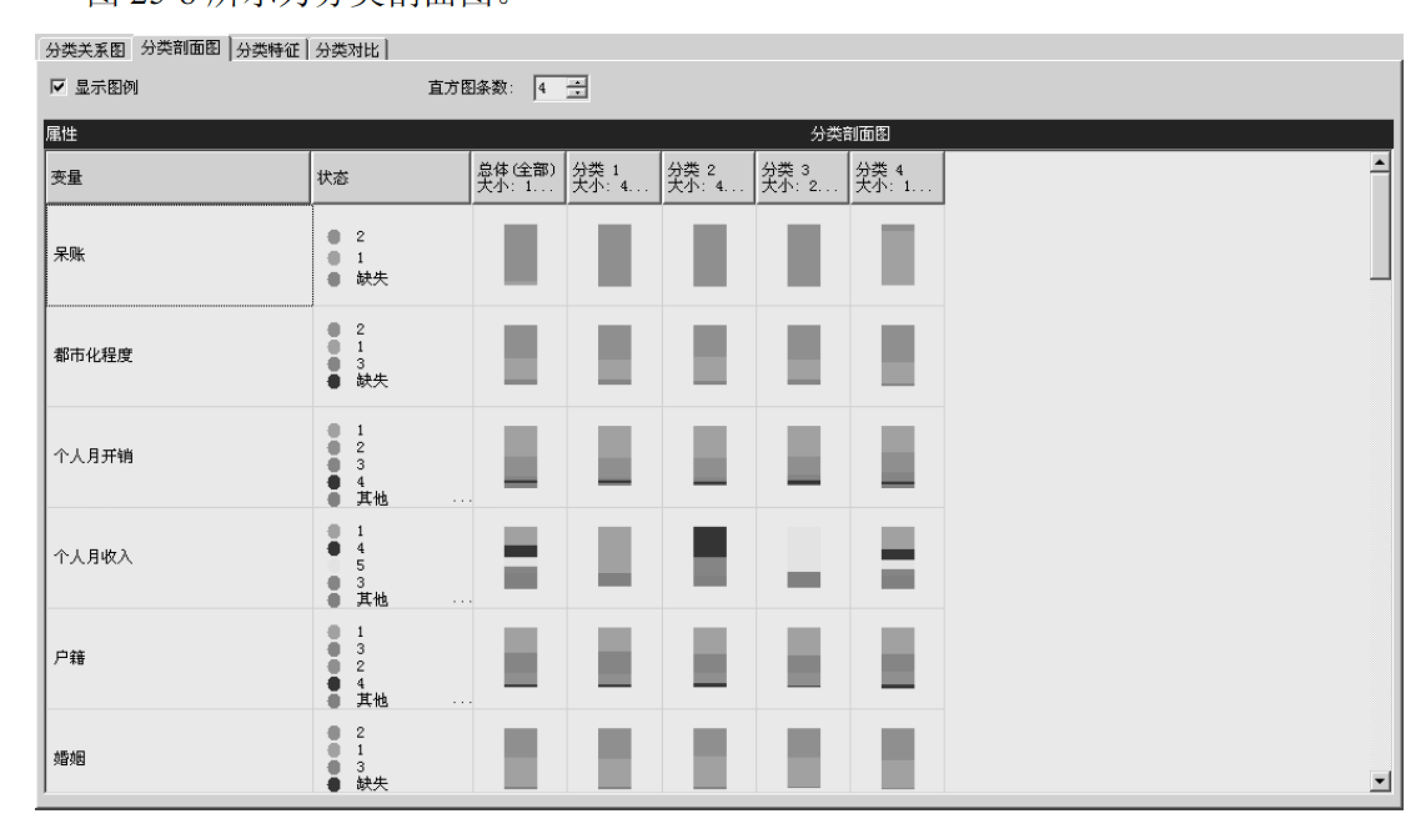


图 25-8　分类剖面图

通过剖面图的分析可以看出，1、2、3 类都不存在呆账情况，4 类呆账比较明显。各类人群在个人月开销中和都市化程度上差别不大，也就是说信用卡的评级好坏可能与这两个变量关系不大。各群的个人月收入之间差异较大，第 1 类收入较低，大部分在 10 000 元以下，第 2 类主要集中在 20 000～40 000 元之间，第 3 类主要在 50 000 元以上，第 4 群各收入阶层的都有。借款余额只有第 4 类是大于 800 万元，其余都没有。拒往记录和强制停卡记录也只是第 4 类有，其余的没有。其余变量之间相差不大。

（1）分类特征图

图 25-9 所示为分类特征图。



图 25-9　分类特征图

从分类特征图可以详细地看出每一群持卡人的特征。通过前面的分析和图 25-9 的对比分析，我们可以看出分类 4 是一个风险性非常高的群，具有强制停卡记录，逾期不还的不良记录，还有退票记录（即支票不能兑现），借款余额超过 800 万元，具有呆账记录（账不能收回），具有瑕疵，而这些分类 1、分类 2 和分类 3 都不具有。根据上面的分析，将分类 4 定为高风险群，建议银行在给该类人群办理信用卡时要慎重，如果情况较为严重，最好不要给该类人群办理信用卡。

（2）分类对比图

通过分类对比图，可以看出群两两之间的差别和对比关系，如图 25-10～图 25-12 所示。在本报告中把关联性最强的分类 2 和分类 3 进行对比，分类 1 和分类 4 进行对比。



图 25-10 分类对比图 1



图 25-11 分类对比图 2

分类 1、分类 2 和分类 3 的关联性不是很强，分类 2 和分类 3 的关联性最强，分类 1 和分类 2 的对比分析如下：分类 1 的家庭月收入较低，集中在 1、2（20 000 元以下和 20 001～40 000 元）附近，收入较低；分类 2 的家庭月收入集中在 3、4（40 001～60 000 元和 60 001～80 000 元），收入中等；分类 3 的家庭月收入集中在 5、6（80 001～100 000 元和 100 001

元以上），属于高收入家庭。分类 1 的个人月收入较低，集中在 1、2（无收入和 10 000 元以下）附近，收入较低；分类 2 的个人月收入集中在 3、4（10 001～20 000 元和 20 001～30 000 元），收入中等；分类 3 的个人月收入集中在 5、6（80 001～100 000 元和 100 001 元以上），属于高收入人群。同时注意到，有部分个人月收入很高，达到 50 001～60 000 元，甚至 60 001 元以上，但是由于他们属于“购物狂”一族，月刷卡额达到 4、5（60 001～80 000 元和 80 001～100 000 元），他们也属于分类 2。



图 25-12　分类对比图 3

通过以上分析，可以得出分类 1、2、3、4 的信用等级分别为低、中、高和极低。分类 1 是较高度风险申请者，对待分类 1 的顾客要注意防范和控制风险，因为他们的收入和家庭经济比较低，如果产生太多的透支，可能导致他们不能按时还款；分类 2 属于中度风险申请者，他们的收入和家庭经济属于中等，这一群人也是数量较多的一类，合理地设定好他们的消费限额，不仅有利于控制他们的风险，而且也能很好地为银行创收。分类 3 是低度风险申请者，这一类人是高收入、家庭经济非常良好的一类，还款能力较强，在申请信用卡时可针对他们提供很多优惠措施，经常为其提供个性化良好服务。分类 4 是高风险申请者，这一群人以前就有过不良的信用卡记录，在对他们发放信用卡时要特别留心，花更多的时间进行审核。

3．精确度分析

（1）借款余额精确度分析

图 25-13 为挖掘结构的资料挖掘精确度借款余额增益图，借款余额的预测概率为 99.21%。

图 25-14 为挖掘结构的资料挖掘精确度借款余额的分类矩阵。

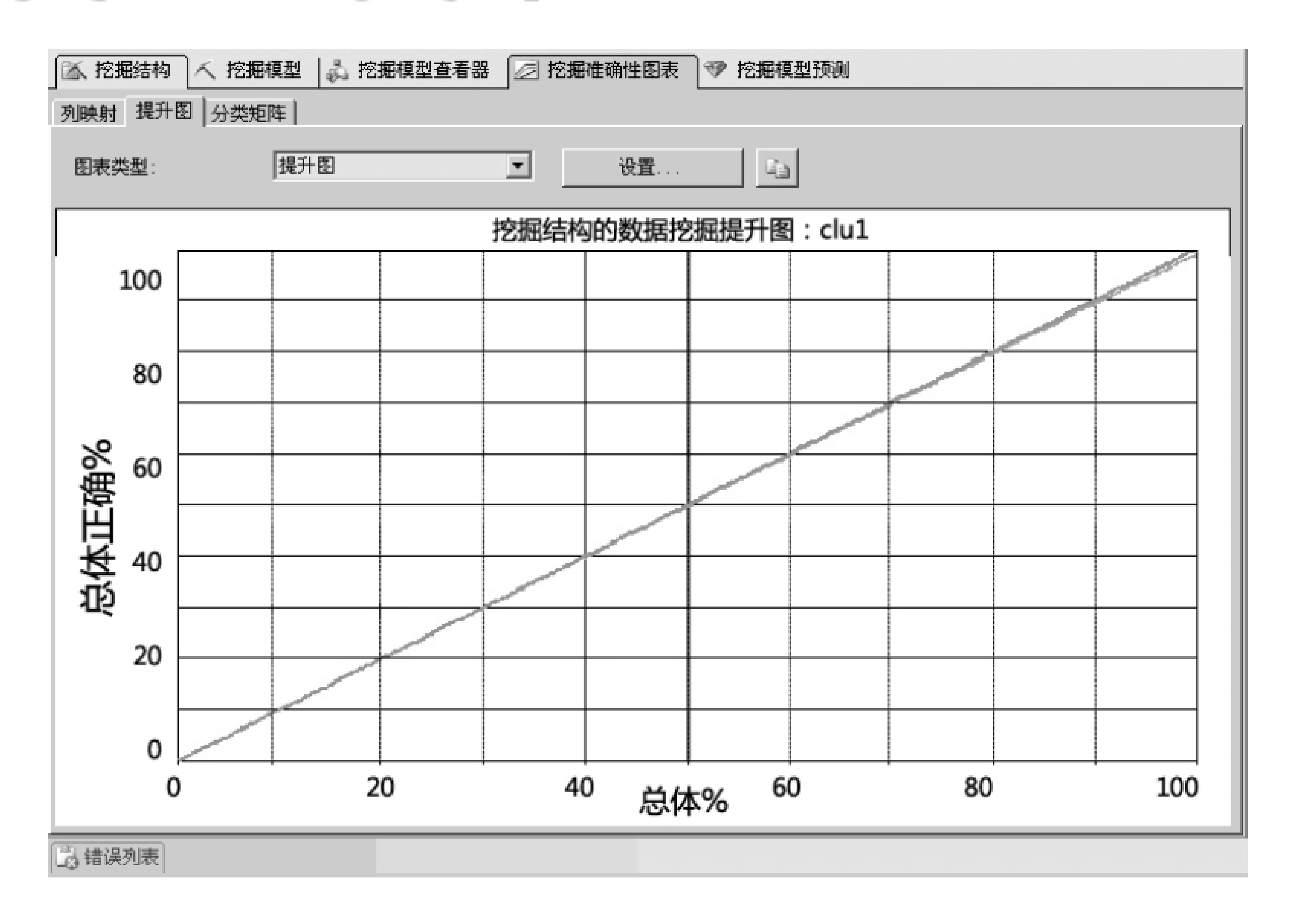
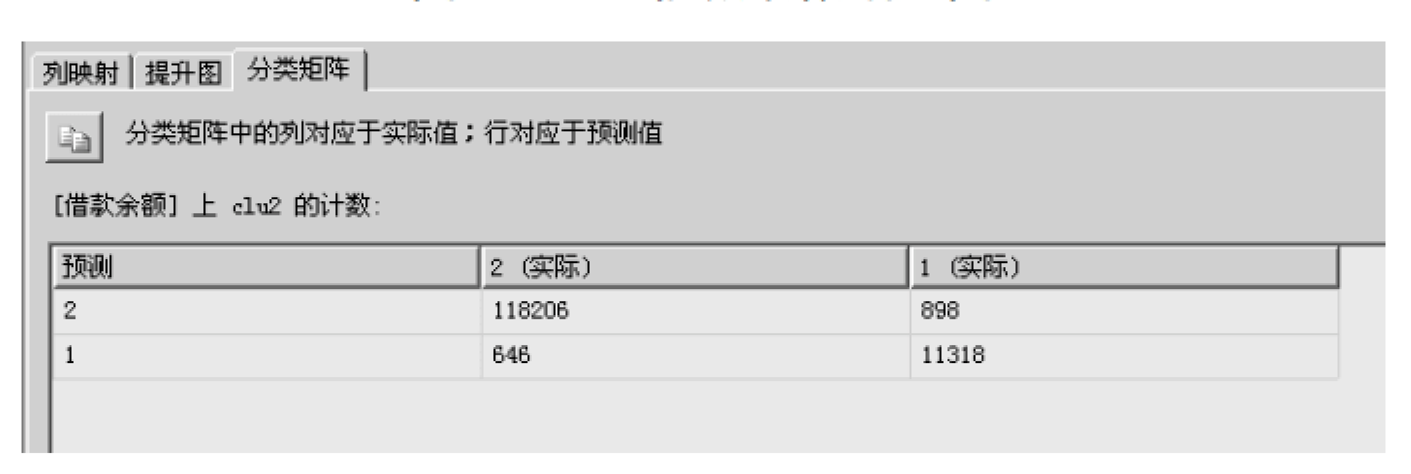


图 25-13　借款余额增益图



列映射 | 提升图 | 分类矩阵

分类矩阵中的列对应于实际值；行对应于预测值

[借款余额] 上 clu2 的计数:

预测	2 (实际)	1 (实际)
2	118206	898
1	646	11318

图 25-14　借款余额分类矩阵

从预测借款余额的角度上讲，这个模型的预测效果很好。

（2）瑕疵户精确度分析

图 25-15 为挖掘结构的资料挖掘精确度瑕疵户增益图，借款余额的预测概率为 100%。

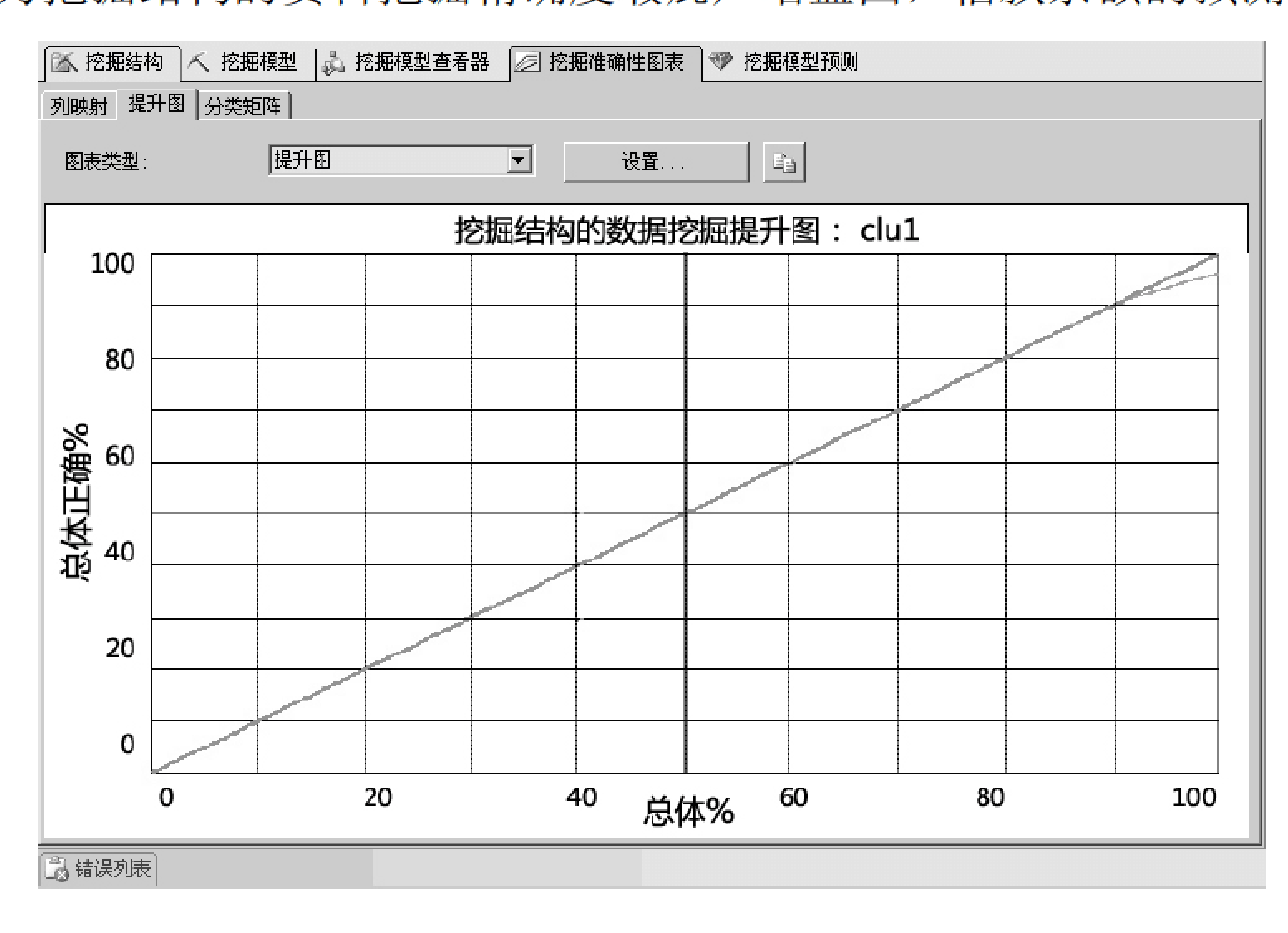


图 25-15　瑕疵户增益图

图 25-16 为挖掘结构的资料挖掘精确度瑕疵户的分类矩阵。

分类矩阵中的列对应于实际值；行对应于预测值

[瑕疵户] 上 clu2 的计数:

预测	2（实际）	1（实际）
2	119104	0
1	4780	7184

图 25-16　瑕疵户分类矩阵

25.4.3　Logistic 回归

1．模型的建立及解释

以借款余额（大于 800 万元，小于 800 万元）为因变量，其他变量为自变量，希望以其他变量去预测借款余额，找出哪些变量与借款余额有关系，如图 25-17 所示。

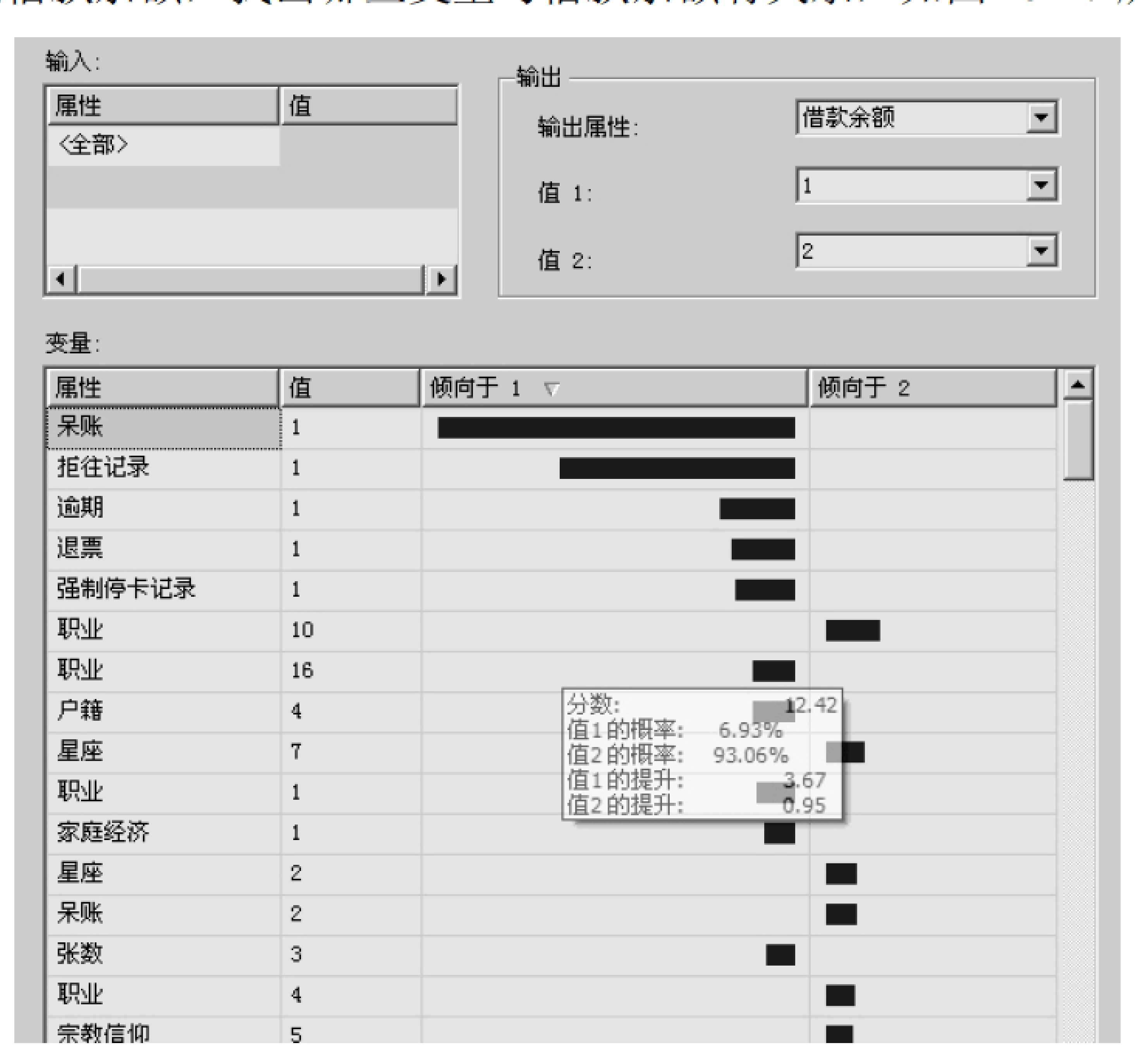


图 25-17　建立模型

借款余额大于 800 万元的 99.89%都具有呆账记录，96.33%有拒往记录（直接被银行拒掉）。

2．模型的准确性分析

从图 25-18 的分析可以看出，模型的预测效果是很好的，达到 99.8%，得分很好，达到了 1，从这一层面来说模型的建立是有效的。

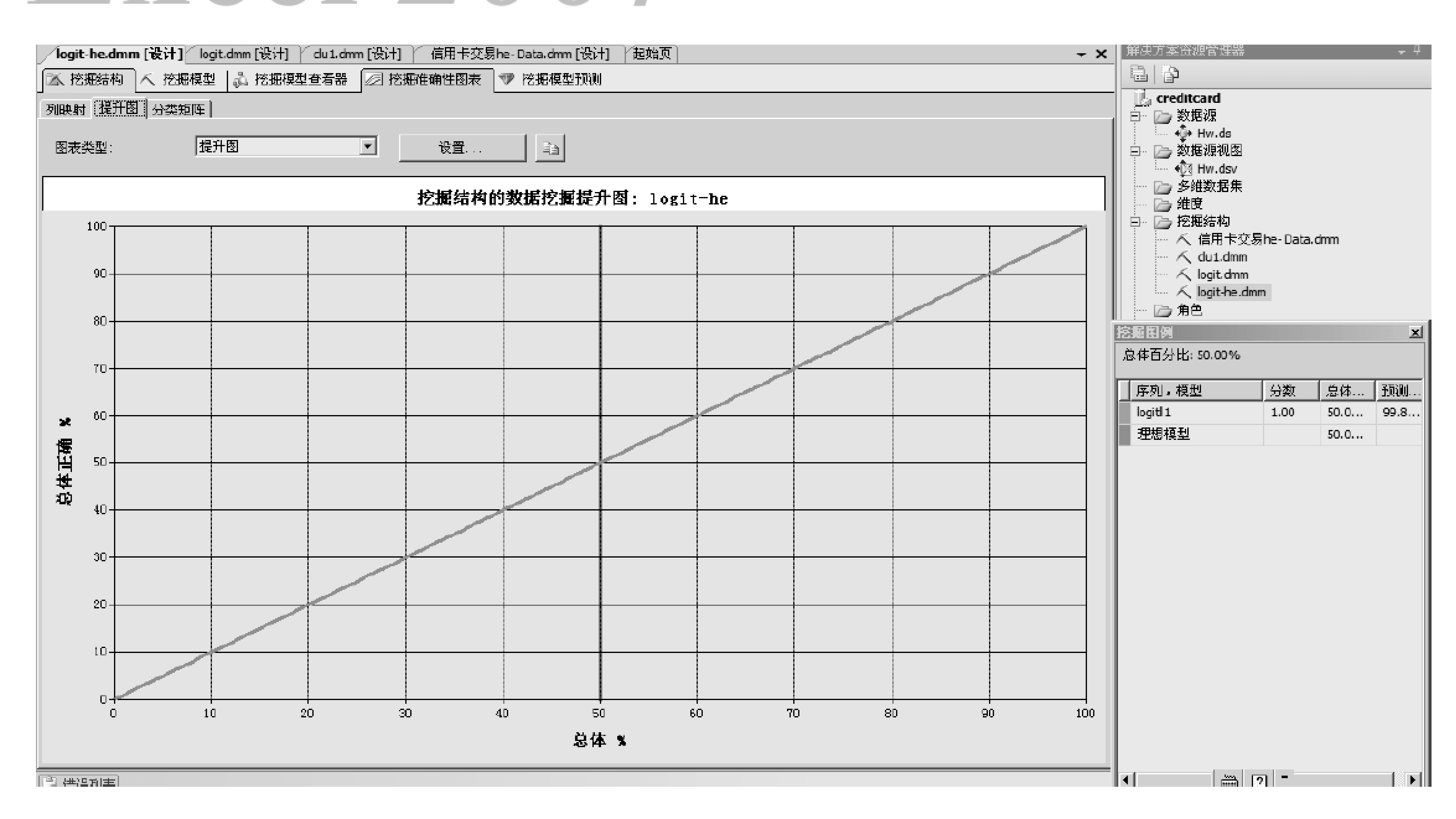
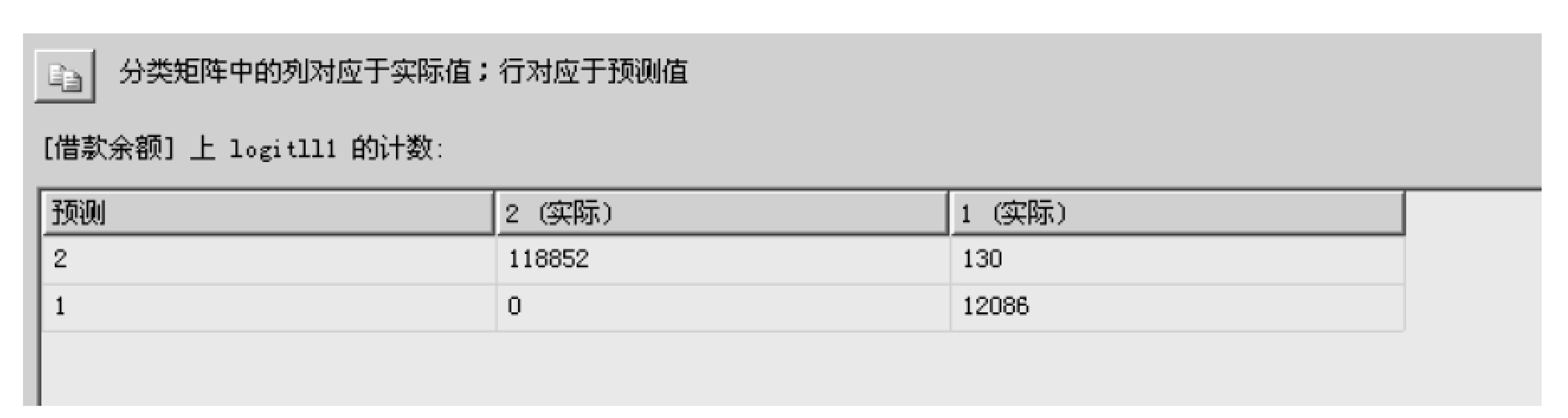


图 25-18　准确性分析

如图 25-19 所示，从分类矩阵看，只有极少数的人借款余额大于 800 万元而预测错了，借款余额小于 800 万元的预测结果都正确。



分类矩阵中的列对应于实际值；行对应于预测值

[借款余额] 上 logitl11 的计数:

预测	2（实际）	1（实际）
2	118852	130
1	0	12086

图 25-19　分类矩阵

第 26 章　市场营销与客户细分的挖掘模型

26.1　研究动机与目的

现阶段，国内的卷烟市场具有垄断与竞争的双重特点。在营销领域，各地烟草公司拥有唯一的卷烟进货权和批发权，通过许可证制度，在卷烟生产、供应、分销和零售等环节建立了严格的专卖专营体系。目前，国内卷烟企业众多，产品差异化程度低，且卷烟生产能力远大于市场需求，导致行业内部竞争空前激烈。随着我国加入 WTO，中国的烟草行业将在未来的几年内面临全面开放的市场，进入国内市场国际化的全面自由竞争局面。中国烟草正迈入一个品牌竞争的时期，我国烟草行业品牌建设与市场营销能力存在很大差距。

因此，积极开展卷烟市场研究，为卷烟生产企业和销售公司品牌培育、营销决策提供参考信息，变得越来越重要。考虑到很多地区的烟草公司都初步建立了包含卷烟零售户的基本资料信息和每月的销售记录的数据库，以数据挖掘的手段来协助市场分析工作就成了一个值得尝试的选择。

26.2　研究方法与限制

相比于其他市场，卷烟销售市场有某些鲜明的特点：按照国内有关法律的限制，卷烟企业品牌展示的主要手段是在零售户的店面内陈列卷烟样品，业内称此为出样。而且卷烟作为一种专卖专营的快速消费品，烟草专卖管理局和烟草销售公司一直有对销售渠道进行管制的权限，体现出“垂直管理、专营专卖”的特点。为了便于管理，管理机构根据零售户在销售时是否遵守相关法规（例如有无经营销售假冒卷烟、乱渠道进烟的行为），又会评出零售户的等级和类别。

现在从数据库筛选出一笔资料：我国北方某省某地区的 1 317 个卷烟零售户在一个月内购买各种品牌卷烟的记录共 91 693 条。而后，依次采用六种分析算法做挖掘分析：决策树、贝叶斯概率、聚类、决策树、Logistic 回归，以及关联分析。

26.3　数 据 分 析

在做挖掘建模之前，先对数据的几项变量作基本的描述统计分析。

先考察这 1 317 个卷烟零售户的结构，表 26-1 反映了这些零售户店主的文化素质构成。

表 26-1　卷烟零售户店主文化素质构成

文 化 素 质	记 录 数 量	比重/%
缺失	39	2.9
初中	744	55.6
大学以上	16	1.2
大专	43	3.2
高中	419	31.3
小学	78	5.8

可见，绝大部分零售户的文化程度为初中和高中。

从表 26-2 可知，大部分客户属于 3 星级的 B 类户，占 57.9%，其次是 4 星级的 A 类户，占 39.3%。

表 26-2　客户分类

	客 户 类 别											
	缺　　失				A 类				B 类			
	记录数	在该列中比重/%	在该行中比重/%	在表中比重%	记录数量	在该列中比重/%	在该行中比重/%	在表中比重/%	记录数量	在该列中比重/%	在该行中比重/%	在表中比重/%
缺失	22	100	100	1.6	0	0	0	0	0	0	0	0
2 星	0	0	0	0	0	0	0	0	16	2.0	100	1.2
3 星	0	0	0	0	0	0	0	0	775	98.0	100	57.9
4 星	0	0	0	0	526	100	100	39.3	0	0	0	0

图 26-1 所示为卷烟零售户每月销售额分布。

图 26-2 所示为卷烟零售户每月销售总数分布。

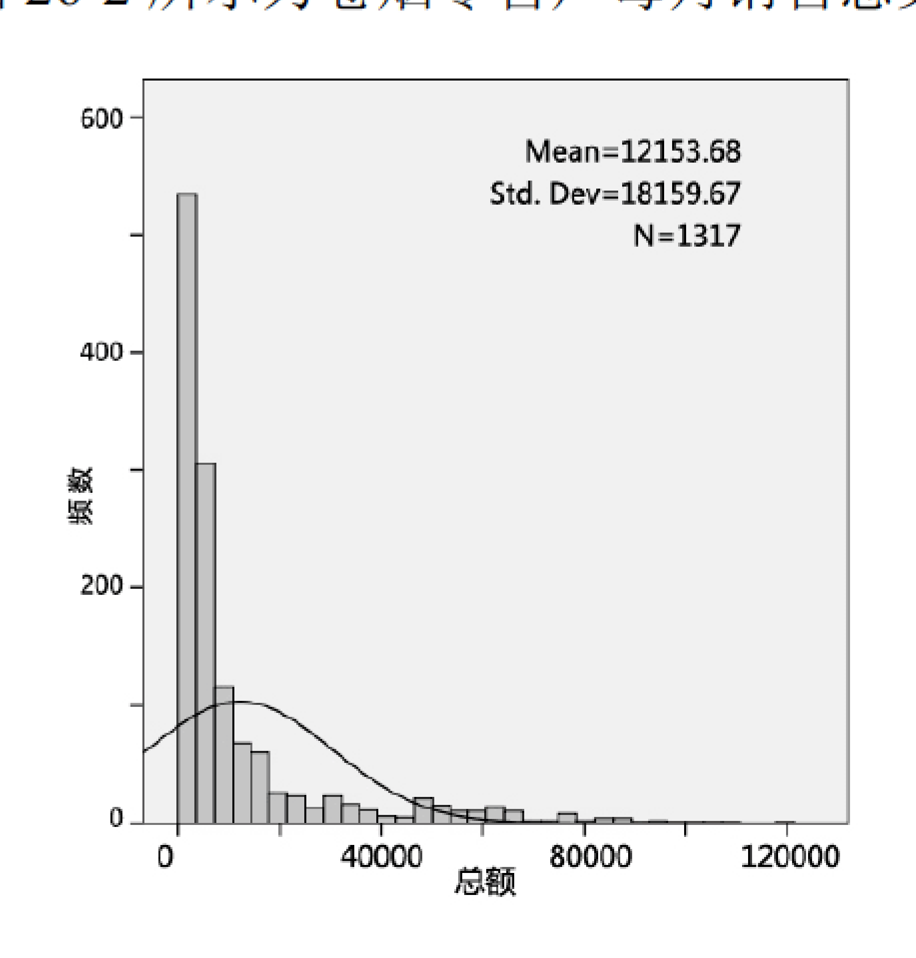


图 26-1　卷烟零售户每月销售额分布

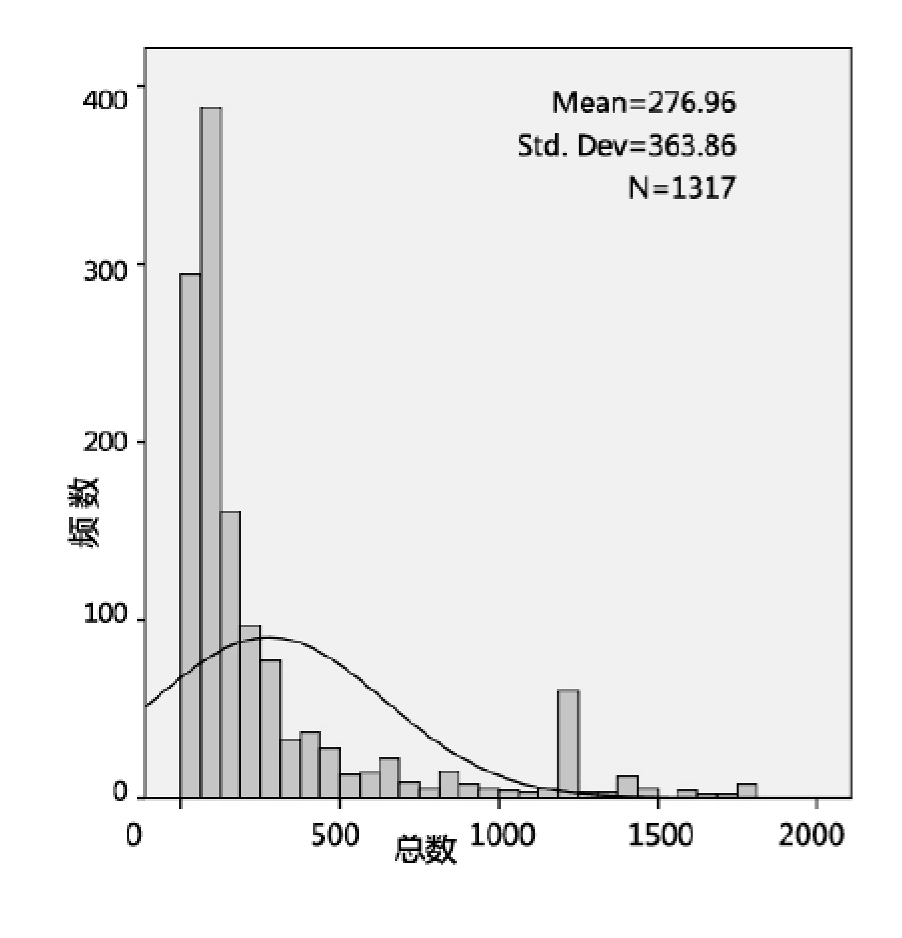


图 26-2　卷烟零售户每月销售总数分布

图 26-3 所示为卷烟零售户利润分布。

分析本月卷烟销售记录的频数，按照零售户的“地段人气”属性，分类如表 26-3 所示。

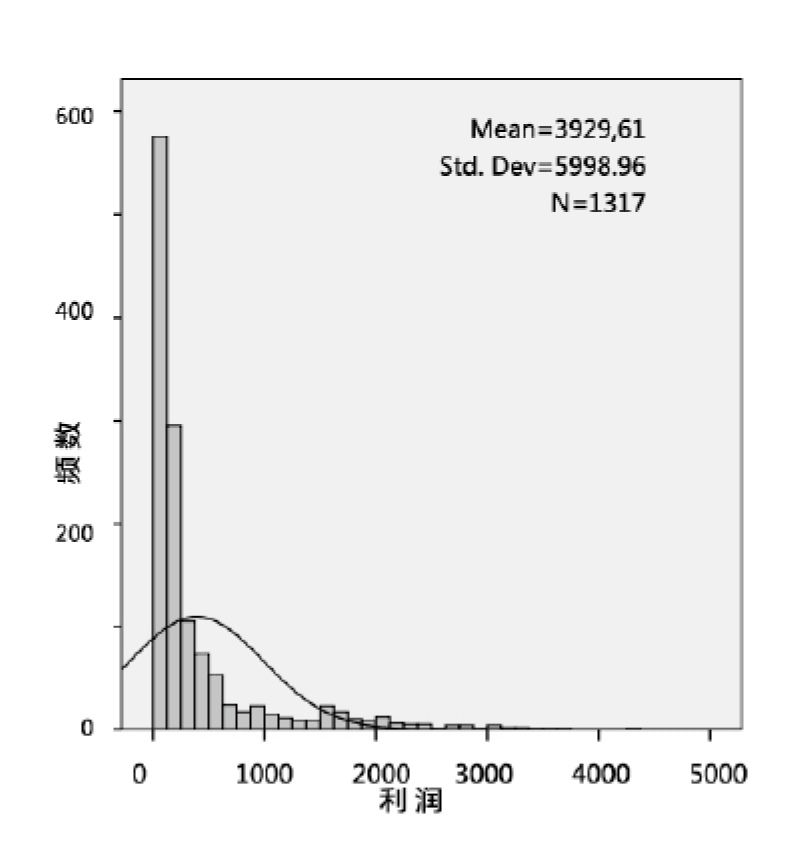


图 26-3　卷烟零售户利润分布

表 26-3　按“地段人气”属性分类

地 段 人 气	频　　数	百分比/%
繁华交通要道	13 312	14.66
较偏远	7 833	8.63
居民区一般街道	67 418	74.27
偏远山区	180	0.20
市内主要商业街	2 034	2.24
总　　计	90 777	100

可见，本月绝大多数的销售事项是发生在店面在居民区一般街道的零售户这一群体中的，其次是店面在繁华交通要道的零售户和较偏远的零售户。

26.4　挖 掘 建 模

26.4.1　决策树

香烟零售市场存在数百个品种的卷烟，一一分析其销售规律工作量过大。不妨参考我国的烟草销售制度，将不同种类香烟按照价格分为五个等级。

一类烟：不含税调拨价 100 元/条以上。

二类烟：不含税调拨价 50～100 元/条。

三类烟：不含税调拨价 30～50 元/条。

四类烟：不含税调拨价 16.5～30 元/条。

五类烟：不含税调拨价 16.5 元/条以下。

从中研究每种等级香烟的销售规律。挖掘模型以自动编码 row_index 为索引键，“香烟等级”为因变量，建立了两个决策树模型，分别命名为 tree 和 tree1。其中模型 tree 的自变量为所有其他变量，模型 tree1 则在自变量列表中删掉了“利润”变量。

表 26-4、表 26-5 为 tree 模型和 tree1 模型的自变量设定。

表 26-4　tree 模型的自变量设定

输入自变量名	含　　义	类　　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
利润率	每种卷烟的零售利润率	连续变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量

表 26-5　tree1 模型的自变量设定

输入自变量名	含　　义	类　　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量

1．数据挖掘模型查看器

（1）模型 tree

图 26-4 为模型 tree 的分类树结构，模型 tree 总共有 6 层，自决策树的顶端向下，前五层的变量都是利润率。关联性强度为：利润率＞总数＞主管部门。由此可知零售户每次购

买的卷烟等级与该品牌卷烟的利润率（即进销价差额除以进价）有最强的关联性，如图 26-5 所示。

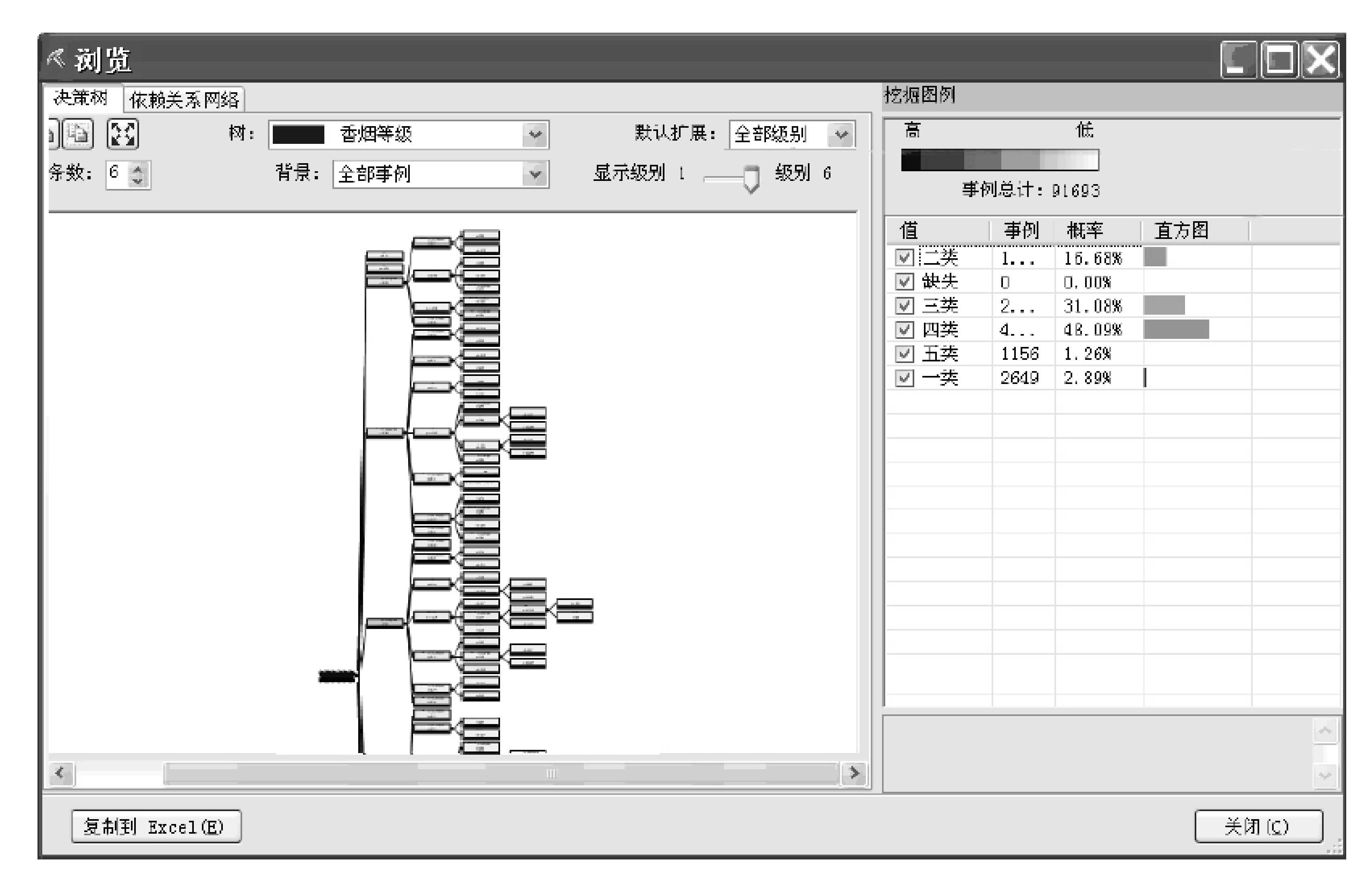


图 26-4　模型 tree 分类树结构

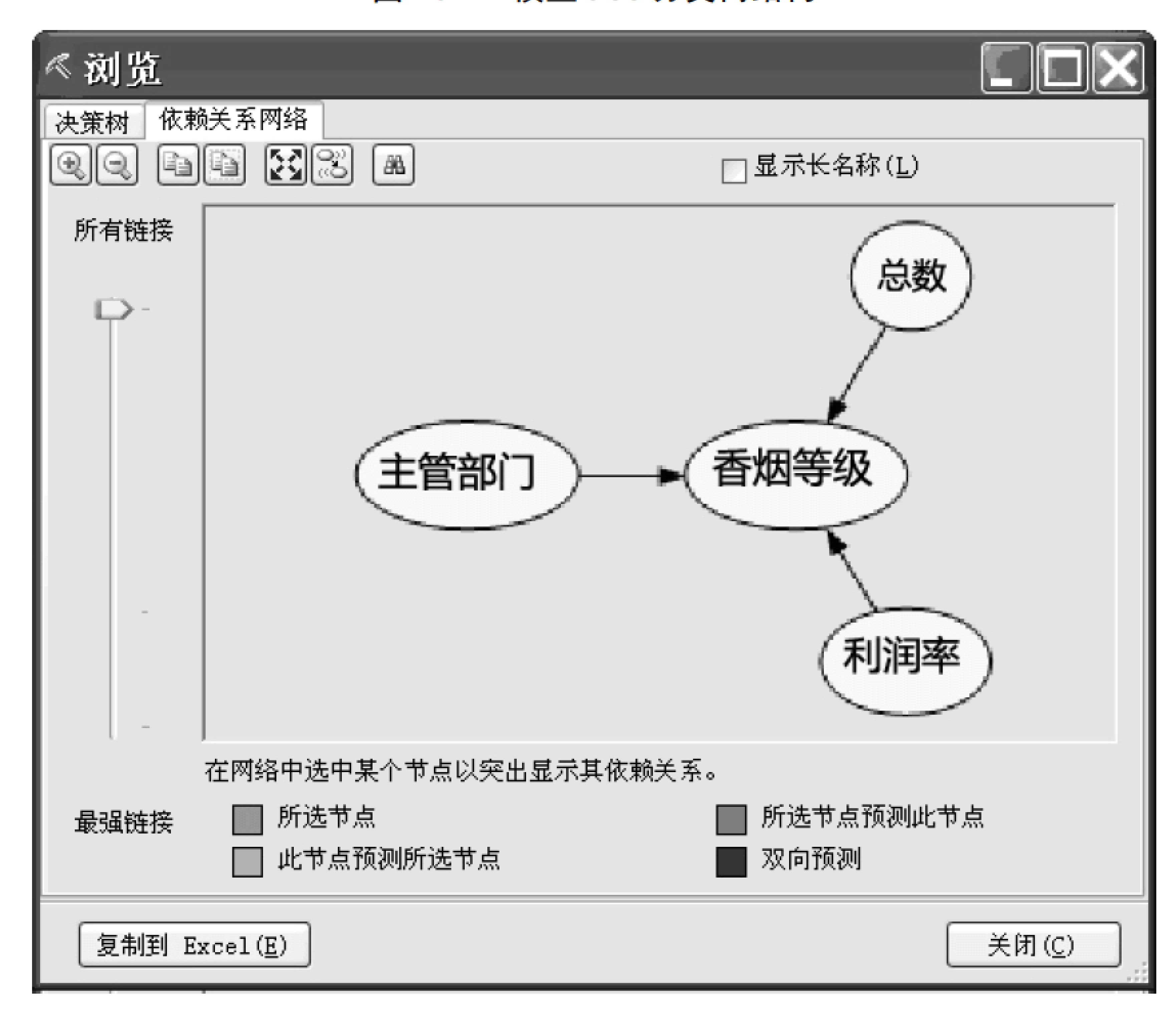


图 26-5　模型 tree 关联性强度图

（2）模型 tree1

模型 tree1 是忽略最强关联性的变量“利润率”后得出的，这样可以更加详细地反映其他自变量和香烟等级的关系。图 26-6 为模型 treel 的分类树结构，模型 tree1 决策树为六层，

第一层是“客户星级”。图 26-7 为 tree1 的关联性强度图，关联性强度大小为：主管部门＞总数＞地段人气＞客户星级＞客户类别＞出样能力，由此可知零售户某次购进的卷烟等级与其月销售量及其“主管部门”有着最强的关联性。



图 26-6　模型 tree1 分类树结构



图 26-7　模型 tree1 关联性强度图

2．数据挖掘精确度图表

为了检验上述两个决策树模型 tree、tree1 的效能，可以采用画准确性图表的方法。例如图 26-8 所示的模型 tree 预测“四类烟”的准确性图表，预测能力均接近理想模型，显示该模型均能做有效的分类和预测。当然，该模型也有其缺点，就是强调“利润率”自变量，

而忽视了其他自变量同香烟等级的关系。

图 26-8　模型 tree 预测“四类烟”准确性图表

而反观模型 tree1，总体来说，其预测准确度比模型 tree 低。例如图 26-9 和图 26-10 所示的预测“四类烟”和“一类烟”的准确性图表。其考虑了“利润率”自变量之外的其他变量与香烟等级的关系，而这些关系的确有重要的意义。例如“卷烟零售主管部门”和“香烟等级”之间的相互影响的关系往往不容易直接观察到。

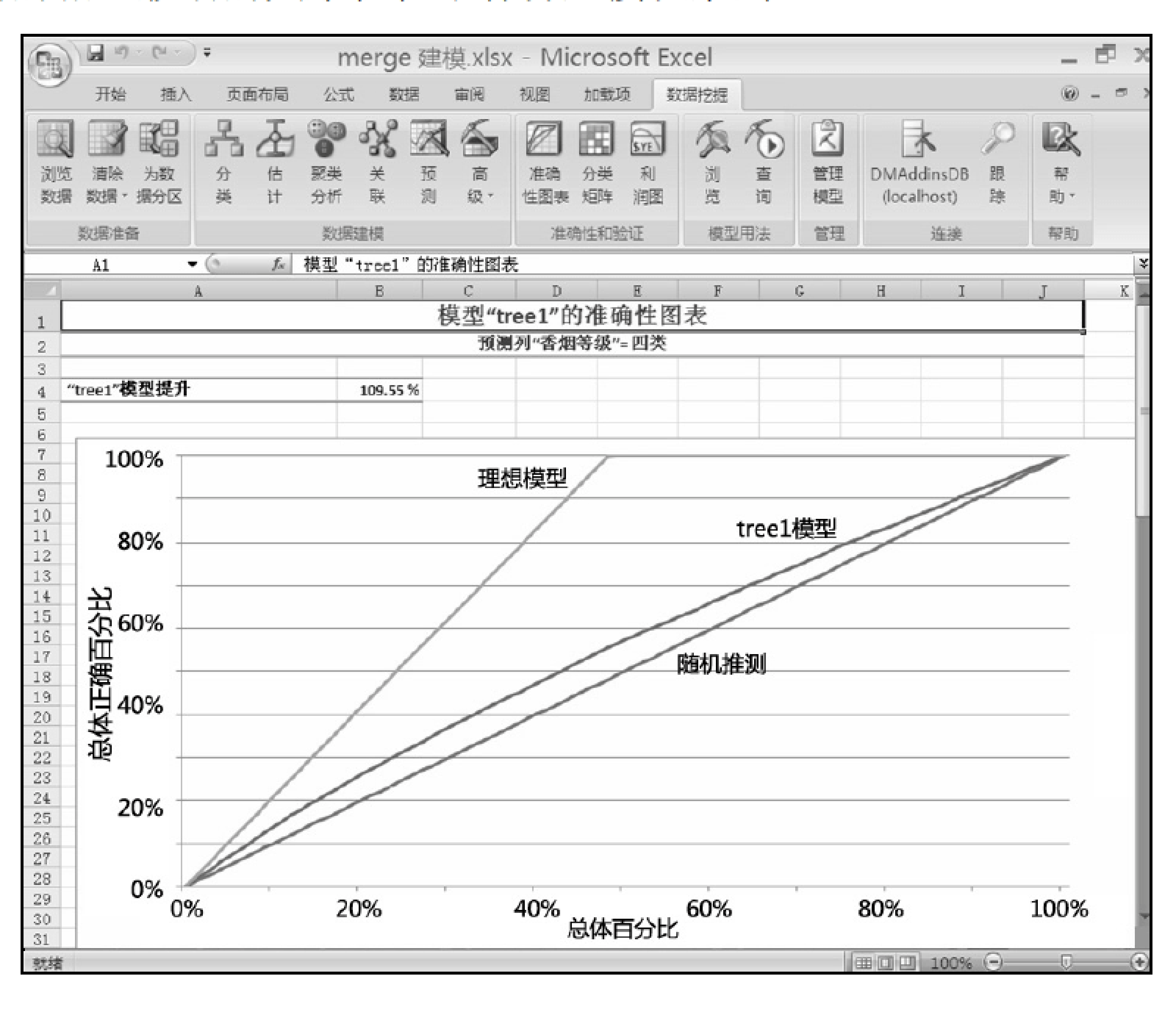


图 26-9　模型 tree1 预测“四类烟”准确性图表

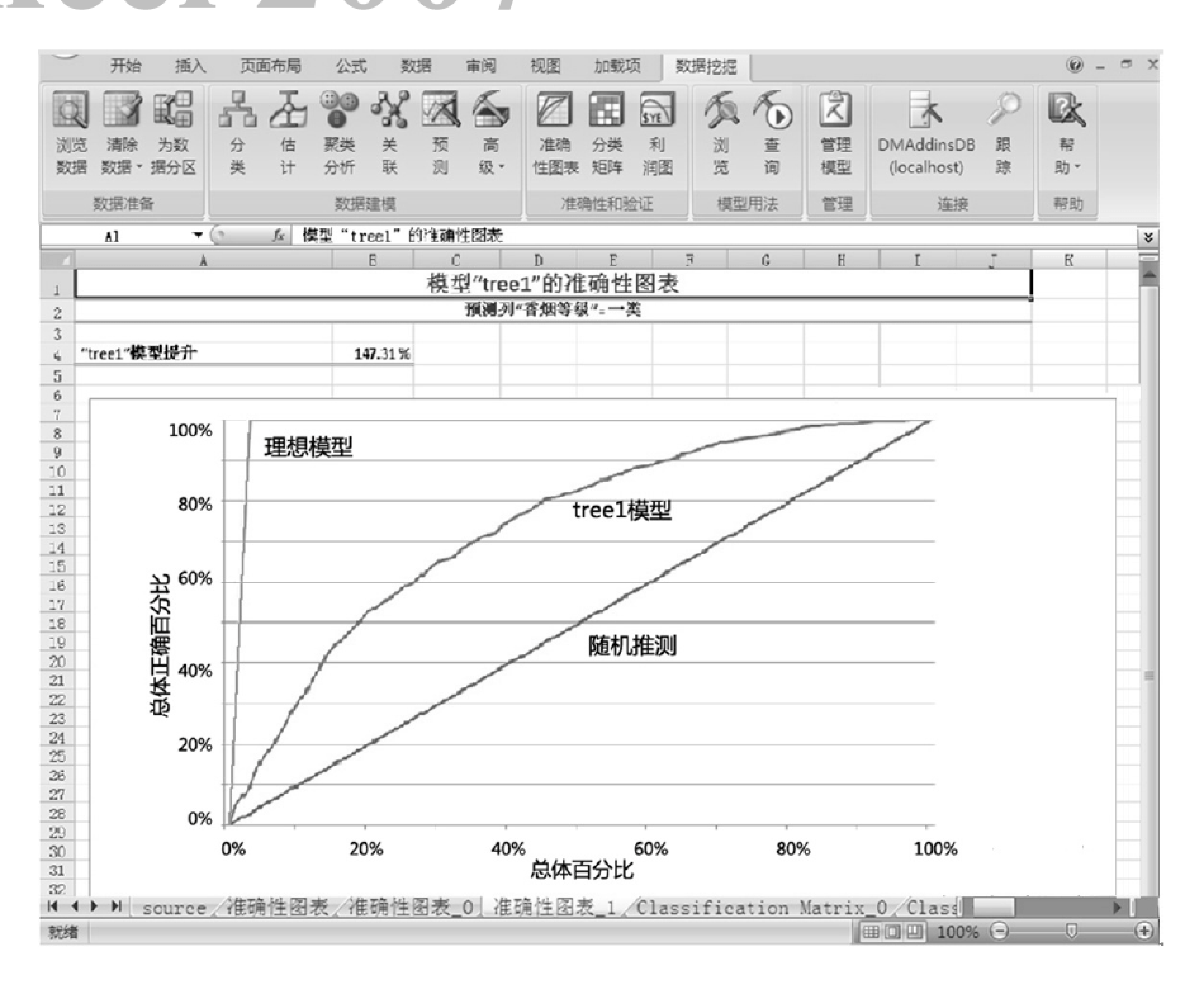
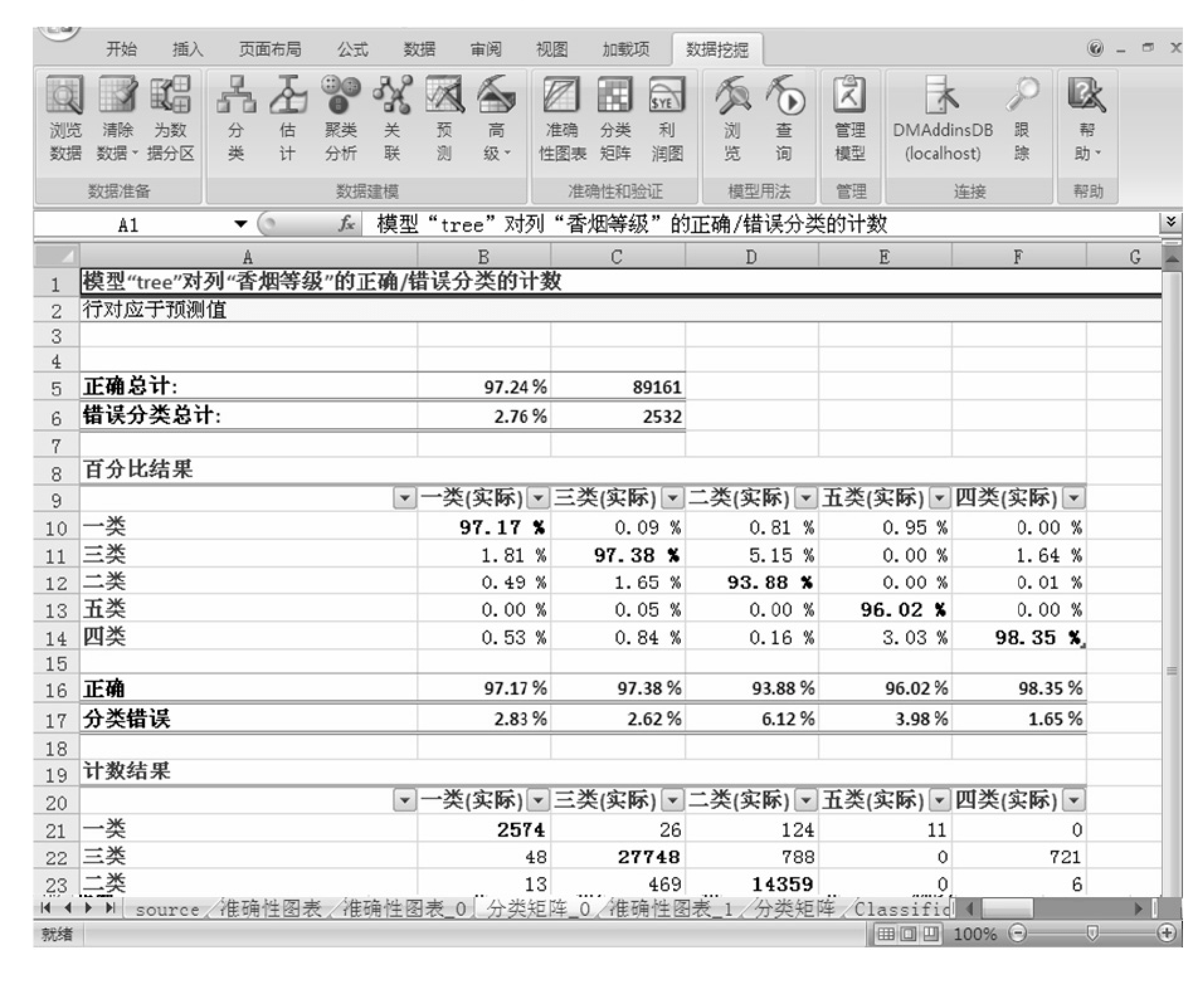


图 26-10　模型 tree1 预测“一类烟”准确性图表

图 26-11 和表 26-6 为 tree 的分类矩阵，由分类矩阵中亦可看出分类的正确率很高。



模型“tree”对列“香烟等级”的正确/错误分类的计数

行对应于预测值

正确总计:	97.24 %	89161
错误分类总计:	2.76 %	2532

百分比结果

	一类(实际)	三类(实际)	二类(实际)	五类(实际)	四类(实际)
一类	97.17 %	0.09 %	0.81 %	0.95 %	0.00 %
三类	1.81 %	97.38 %	5.15 %	0.00 %	1.64 %
二类	0.49 %	1.65 %	93.88 %	0.00 %	0.01 %
五类	0.00 %	0.05 %	0.00 %	96.02 %	0.00 %
四类	0.53 %	0.84 %	0.16 %	3.03 %	98.35 %
正确	97.17 %	97.38 %	93.88 %	96.02 %	98.35 %
分类错误	2.83 %	2.62 %	6.12 %	3.98 %	1.65 %

计数结果

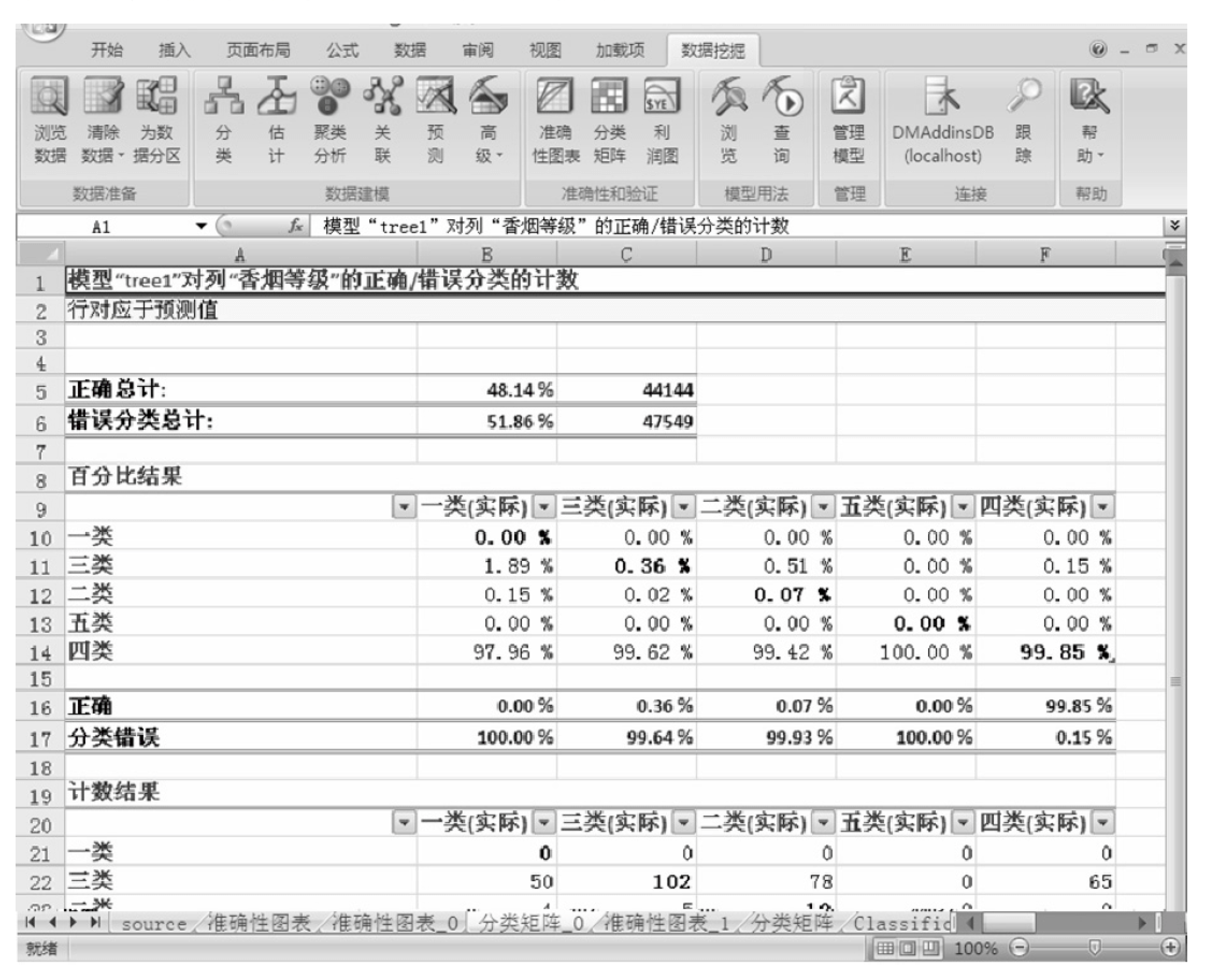
	一类(实际)	三类(实际)	二类(实际)	五类(实际)	四类(实际)
一类	2574	26	124	11	0
三类	48	27748	788	0	721
二类	13	469	14359	0	6

图 26-11　模型 tree 分类矩阵

图 26-6　tree 模型分类矩阵　　　%

	一类（实际）	三类（实际）	二类（实际）	五类（实际）	四类（实际）
一类	97.17	0.09	0.81	0.95	0.00
三类	1.81	97.38	5.15	0.00	1.64
二类	0.49	1.65	93.88	0.00	0.01
五类	0.00	0.05	0.00	96.02	0.00
四类	0.53	0.84	0.16	3.03	98.35
正确	97.17	97.38	93.88	96.02	98.35
分类错误	2.83	2.62	6.12	3.98	1.65

而 tree1 的分类效果并不稳定。针对“四类烟”很高（当然，这非常重要，因为“四类烟”是销售量最大的烟种，给零售户带来的利润仅次于“三类烟”），但是对于其他等级的分类正确率却较低，如图 26-12 和表 26-7 所示。



模型“tree1”对列“香烟等级”的正确/错误分类的计数

行对应于预测值

正确总计:	48.14 %	44144
错误分类总计:	51.86 %	47549

百分比结果

	一类(实际)	三类(实际)	二类(实际)	五类(实际)	四类(实际)
一类	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
三类	1.89 %	0.36 %	0.51 %	0.00 %	0.15 %
二类	0.15 %	0.02 %	0.07 %	0.00 %	0.00 %
五类	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %
四类	97.96 %	99.62 %	99.42 %	100.00 %	99.85 %
正确	0.00 %	0.36 %	0.07 %	0.00 %	99.85 %
分类错误	100.00 %	99.64 %	99.93 %	100.00 %	0.15 %

计数结果

	一类(实际)	三类(实际)	二类(实际)	五类(实际)	四类(实际)
一类	0	0	0	0	0
三类	50	102	78	0	65

图 26-12　模型 treel 分类矩阵

表 26-7　模型 tree1 分类矩阵　　%

	一类（实际）	三类（实际）	二类（实际）	五类（实际）	四类（实际）
一类	0.00	0.00	0.00	0.00	0.00
三类	1.89	0.36	0.51	0.00	0.15
二类	0.15	0.02	0.07	0.00	0.00
五类	0.00	0.00	0.00	0.00	0.00
四类	97.96	99.62	99.42	100.00	99.85
正确	0.00	0.36	0.07	0.00	99.85
分类错误	100.00	99.64	99.93	100.00	0.15

26.4.2　单纯贝叶斯分类

挖掘模型是以自动编码 row_index 为索引键，以香烟的商品品牌为因变量，建立贝叶斯模型设定所有零售户的基本数据为自变量，分析某一款香烟被购进的潜在原因，如表 26-8 所示。

图 26-8　贝叶斯模型的自变量设定

输入自变量名	含　义	类　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量

贝叶斯分析中，图 26-13 为关联性连结图，关联性强度大小为：总数＞客户星级＞客户类别＞主管部门＞出样能力＞是否主营＞结算方式＞营业面积＞文化素质＞地段人气＞订货类型，在此模型中可得出购买何种品牌香烟与商户月销售总数有最强的关联性。

图 26-14 表现了购入某品牌的零售户的各种特征。

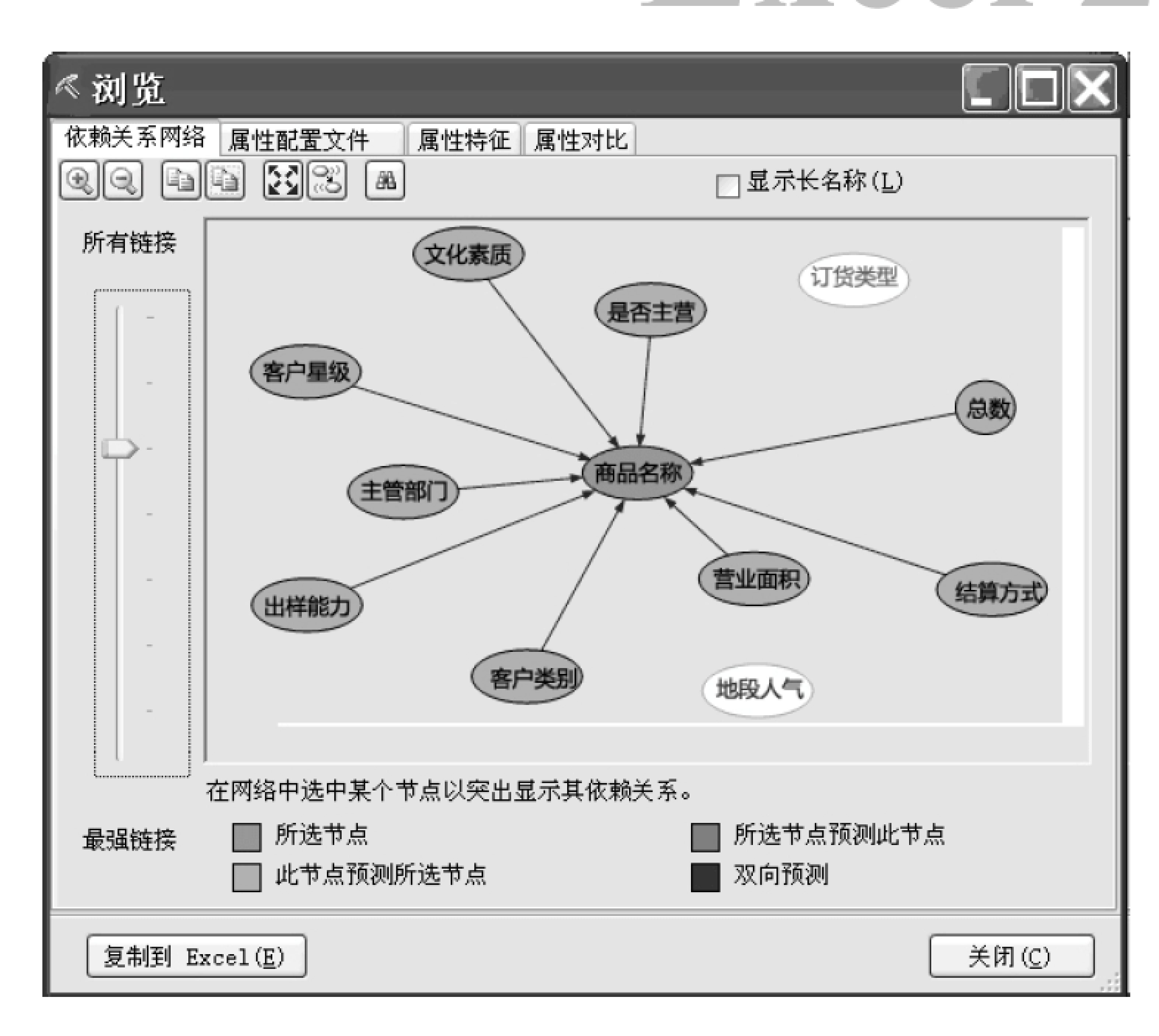


图 26-13　贝叶斯分析分类矩阵



图 26-14　贝叶斯分析属性特征

图 26-15 是在不同自变量条件下，零售户倾向于购买品牌的分析表。

如果对某种商品感兴趣，可以画出该贝叶斯分析模型对此品牌的准确性图表，观察其预测效能。例如图 26-16，如果关心“醇盖红梅”的销售状况，可以画出其准确性图表。发现贝叶斯分析模型在预测商品“醇盖红梅”被零售商采购的预测效能，显著好于随机猜测，但是与理想模型相比，还存在较大的差距。

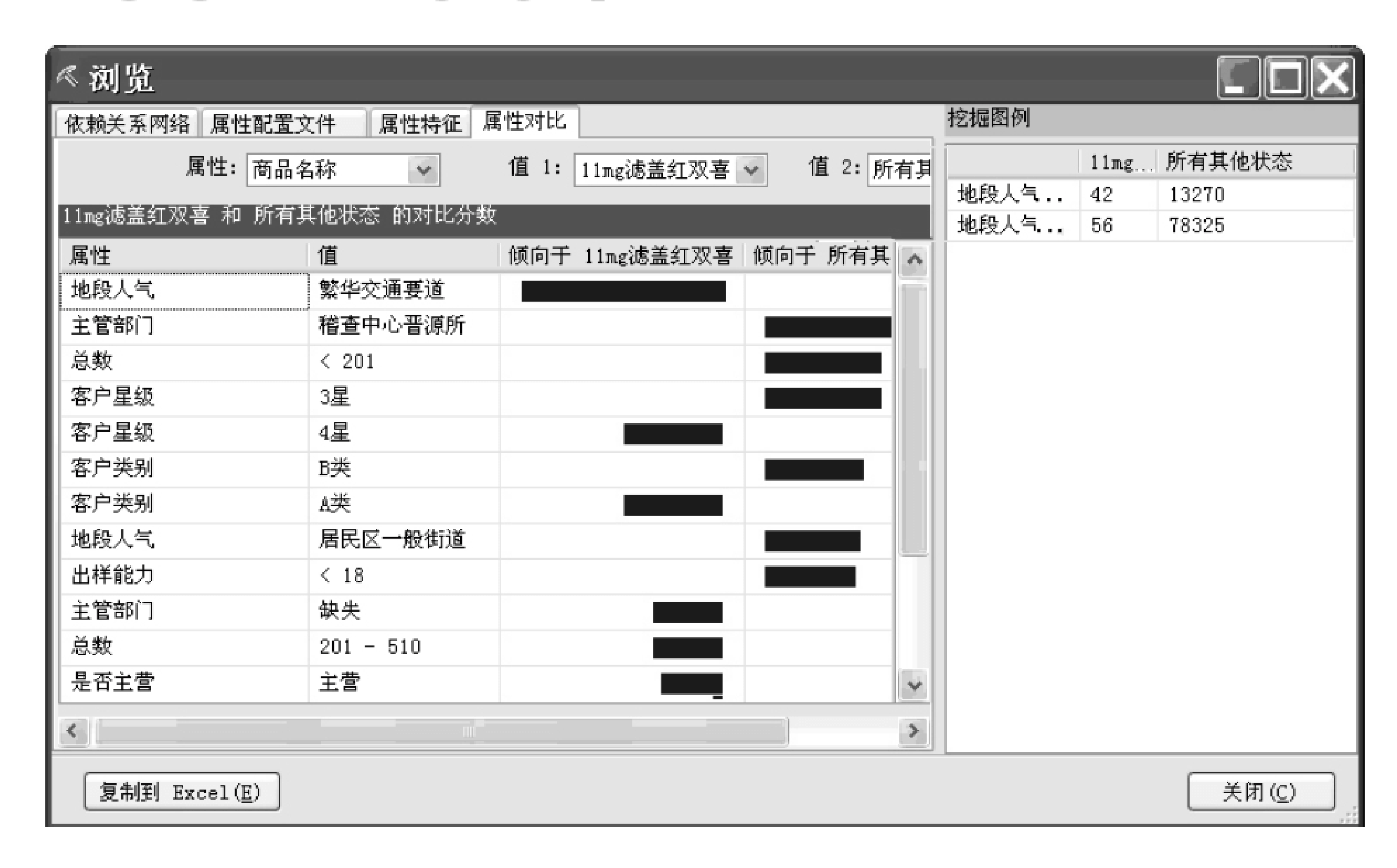


图 26-15 贝叶斯分析属性辨识

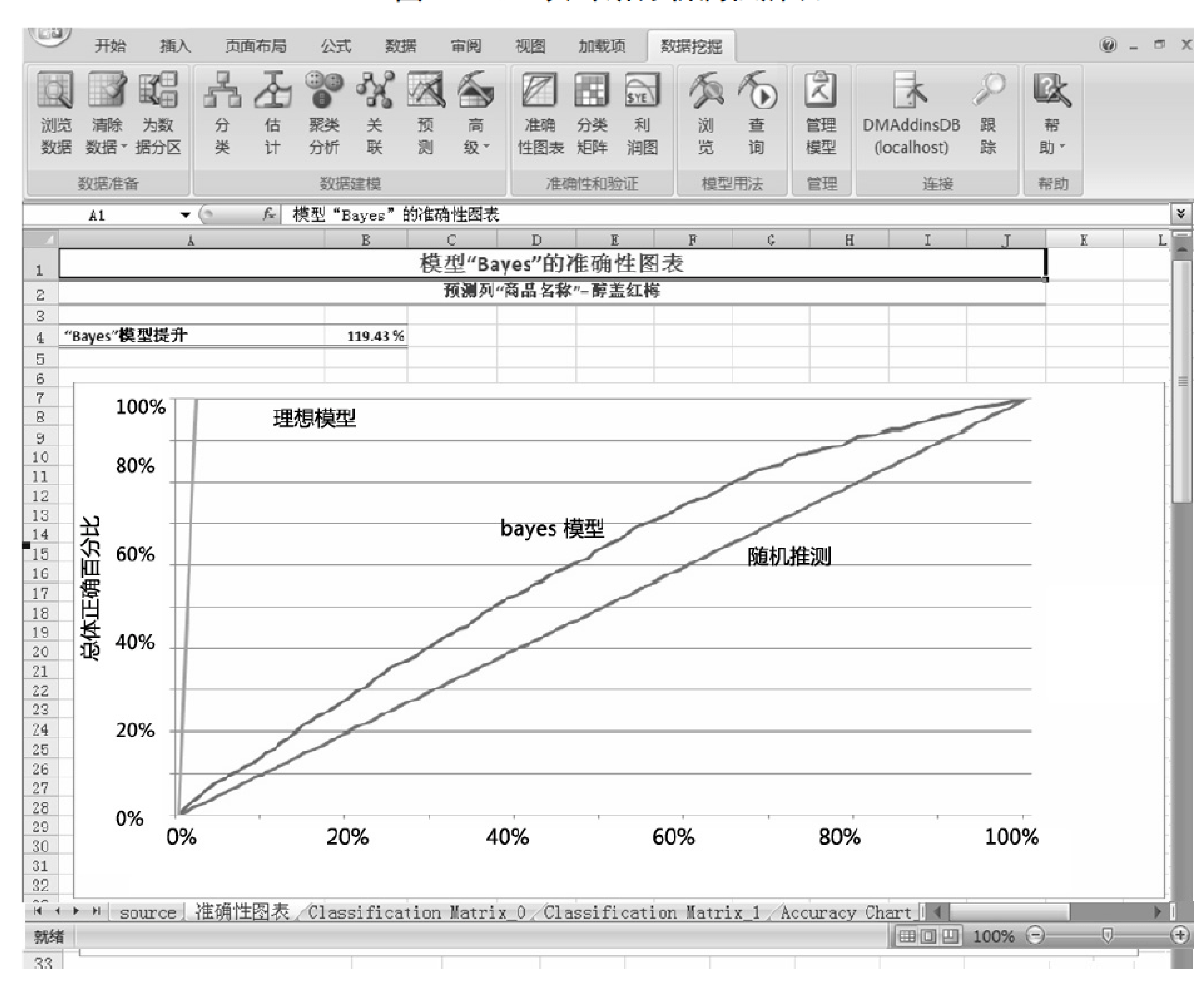


图 26-16 贝叶斯分析准确性图表

26.4.3 聚类分析

构建聚类分析挖掘模型以自动编号 row_index 为索引键，将所有零售户基本资料变量

设为自变量，如表 26-9 所示。按照默认设置，将卷烟零售户集聚成 10 类。

表 26-9　聚类分析的自变量设定

输入自变量名	含　义	类　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量
总额	某商户的总销售金额	连续变量
利润	某商户月卷烟零售利润	连续变量

分类 1 全部采取电话订货的方式，客户类别全部是 B 类户，绝大部分（概率在 90%以上）的零售户持有普通零售的卷烟销售许可证，客户星级为 3 星，月销售额在 125.5～12 153.7 元，月销售利润在 37.8～3 929.6 元，经营形式是兼营，营业面积 0.0～64.4 平方米，月总销售数量为 32～276 条。

分类 2 全部采取电话订货的方式，绝大部分（概率超过 90%）的零售户持有普通零售的卷烟销售许可证，结算方式为电子结算，月销售利润在 37.8～3 929.6 元，月销售额在 125.5～12 153.7 元，客户星级为 4 星，客户类别全部是 A 类户，经营形式是兼营，营业面积为 0.0～64.4 平方米，月总销售数量为 32～276 条。

分类 3 全部采用电话订货的方式，绝大部分（概率超过 99%）为 B 类户，许可证种类为普通零售，客户星级为 3 星，月销售利润在 37.8～3 929.6 元，月销售额在 125.5～12 153.7 元，采用兼营的形式，月总销售数量为 32～276 条。

分类 4 全部采用电话订货，许可证种类全部为普通零售。有很大部分（概率低于 90%，高于 80%）营业户是营业面积为 0.0～64.4 平方米，结算方式为电子结算。大部分（概率低于 80%，高于 70%）月销售额在 24 402.2～66 632.7 元之间，利润在 7 975.9～21 926.5 元之间。

分类 5 的订货方式全部为电话订货，客户星级全部为 4 星，客户类别全部为 A 级，结算方式全部为电子结算，许可证种类全部为普通零售，有绝大部分（概率为 96.88%）营业面积为 0～64.4 平方米。

分类 6 的订货方式全部为电话订货，许可证种类全部为普通零售，有绝大部分（概率

为 92.26%）营业面积为 0.0～64.4 平方米，绝大部分（概率为 90.93%）经营方式为兼营，很大部分（概率低于 90%，高于 80%）结算方式为电子结算，客户星级为 3 星客户类别为 B 类，大部分的（概率低于 80%，高于 70%）商户的“地段人气”类型为居民一般街道，商户法人代表的文化素质为初中。

分类 7 的订货方式全部为电话订货，许可证种类绝大部分（99.99%）为普通零售，绝大部分的（概率为 96.38%）零售户月销售总额为 125.5～12 153.7 元，月销售利润在 37.8～3 929.6 元，并且 88.58%的商户月销售量为 32～276 条。

分类 8 绝大部分（概率在 90%以上）的出样形式为混合型，结算方式为电子结算，许可证为普通零售，订货方式为电话订货。很大部分（概率低于 90%，高于 80%）的商户为 4 星级，A 类户。大部分（59%以上）月销售量为 523～1 368 条，经营方式为主营，店主文化素质为高中。

分类 9 的订货方式全部为电话订货，从业人数全部为 0，意味着店主是自我雇用，相当大比例的零售户有多项资料缺失：例如入网日期、地段人气、主管部门、许可证种类、主营与否、出样形式、文化素质等。94%的商户为 B 类，客户星级为 3 星，营业面积为 0.0～64.4 平方米，但是出样能力为 0 意味着由于经营场地的限制，没有展示品牌的柜台货架，因而没有出样能力。

分类 10 的客户类别全部为 B 类，订货类型为电话订货，绝大部分（概率高于 90%）为兼营的形式。大部分（概率低于 90%，高于 80%）为电子结算。大部分商户的出样形式为混合，客户星级为 2 星，地段的人气大部分也很旺，处于繁华交通要道。

图 26-17 展示了在这些分类之中，零售户的特点。



图 26-17　聚类分析模型聚类概况

图 26-18 为各个分类之间的关联强度示意图。关联强度越强的两类，说明两类之间越是存在明显的规律。

显然，分类 3 和分类 6 为最强的连接。图 26-19 和图 26-20 显示了分类 3 和分类 6 的特征。

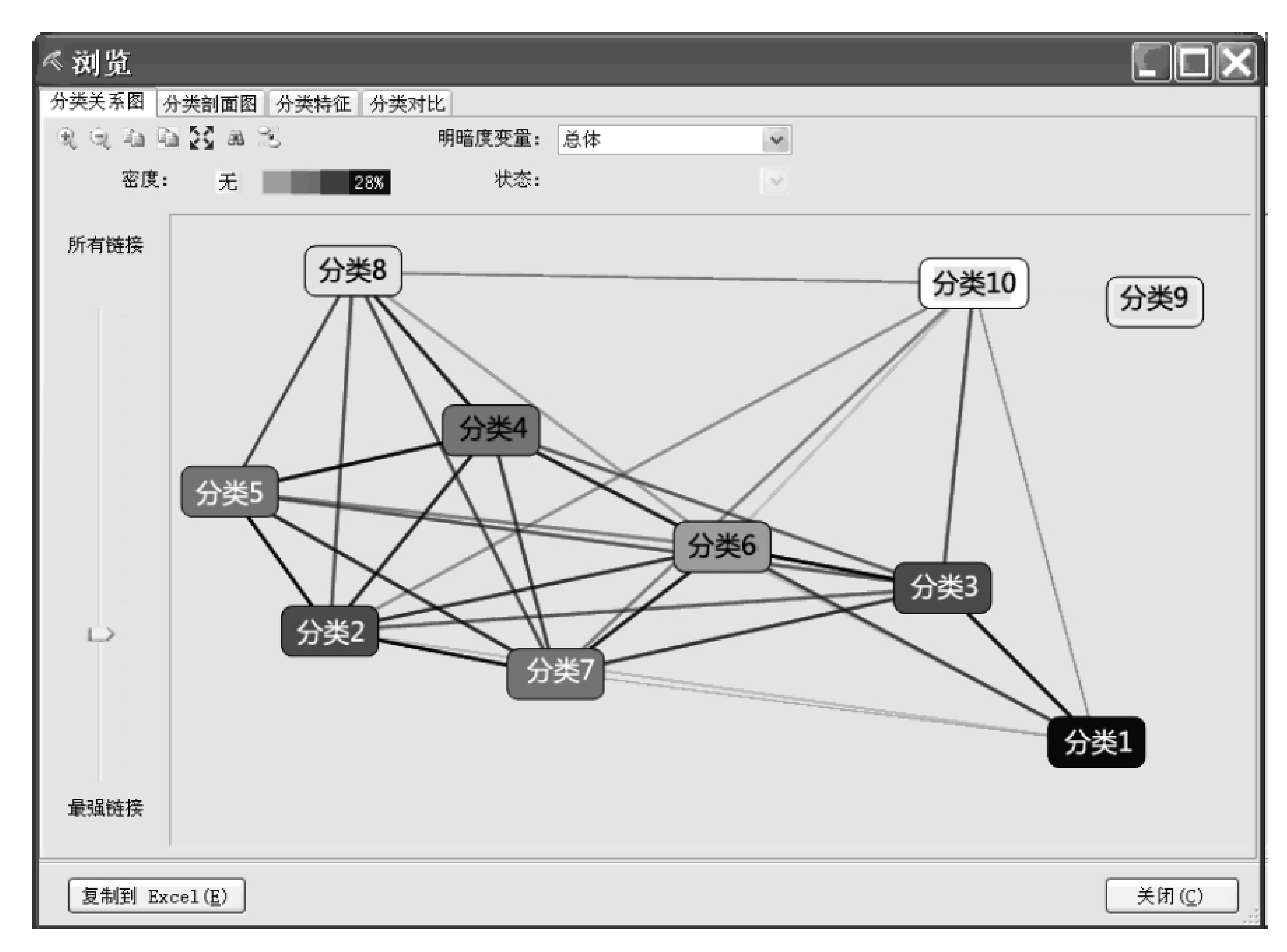


图 26-18　聚类分析模型聚类关联强度



图 26-19　聚类 3 的特征

图 26-21 反映了分类 3 和分类 6 的区别，由此可以找出这两类之间的明显规律性。分

类 3 和分类 6 的月度销售量、销售额和利润都在不同的区间内的可能性非常大。前者的销售分类 3 的上述各项指标比分类 6 偏低。



图 26-20　分类 6 的特征



图 26-21　分类 3 和分类 6 的区别对比

26.4.4　决策树

以零售户利润为因变量，其他零售户资料变量为自变量，建立决策树模型 reg tree1，如表 26-10 所示。决策树模型和多元线性回归模型都可以用来分析多个自变量对于一个变量的影响。之所以考虑决策树模型而非多元线性回归模型，是因为自变量中含有大量的分类变量，而且决策树有助于反映变量之间的非线性关系。在借助 excel-addin 调用 SQL 算法

的时候，只需选择决策树算法，由于因变量是连续变量，所以系统会自动构建决策树。图 26-22 呈现了决策树的结果。

表 26-10　决策树模型的自变量设定

输入自变量名	含　　义	类　　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量
利润	某商户月卷烟零售利润	连续变量

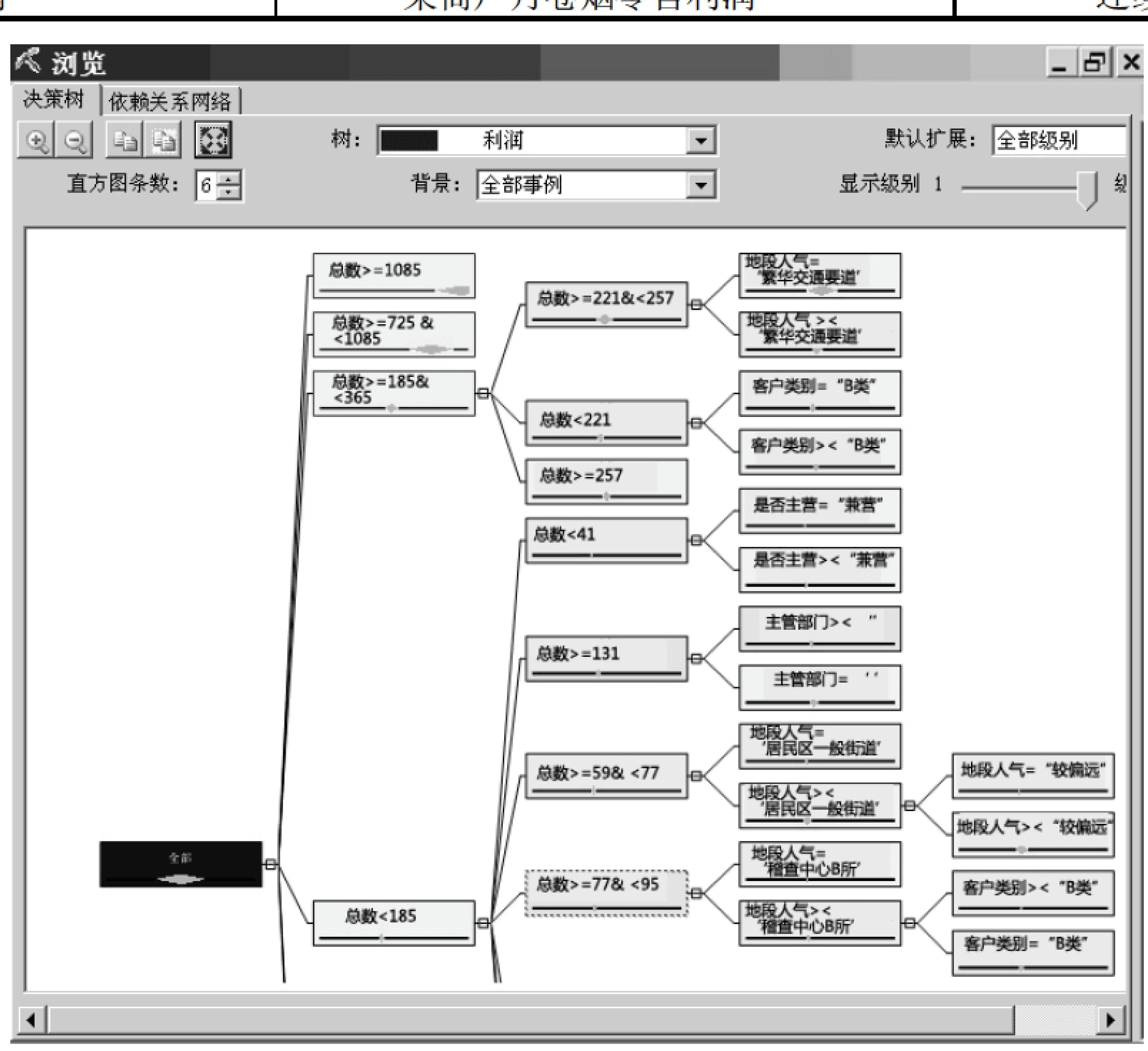


图 26-22　决策树 1 的结果

依赖关系网络则显示了各种自变量对于利润的影响的相对强弱程度，如图 26-23 所示。可见，总数对利润影响最大，其他相关的变量以相依性从大到小排序：地段人气、客户星级、主管部门客户类别、是否主营。

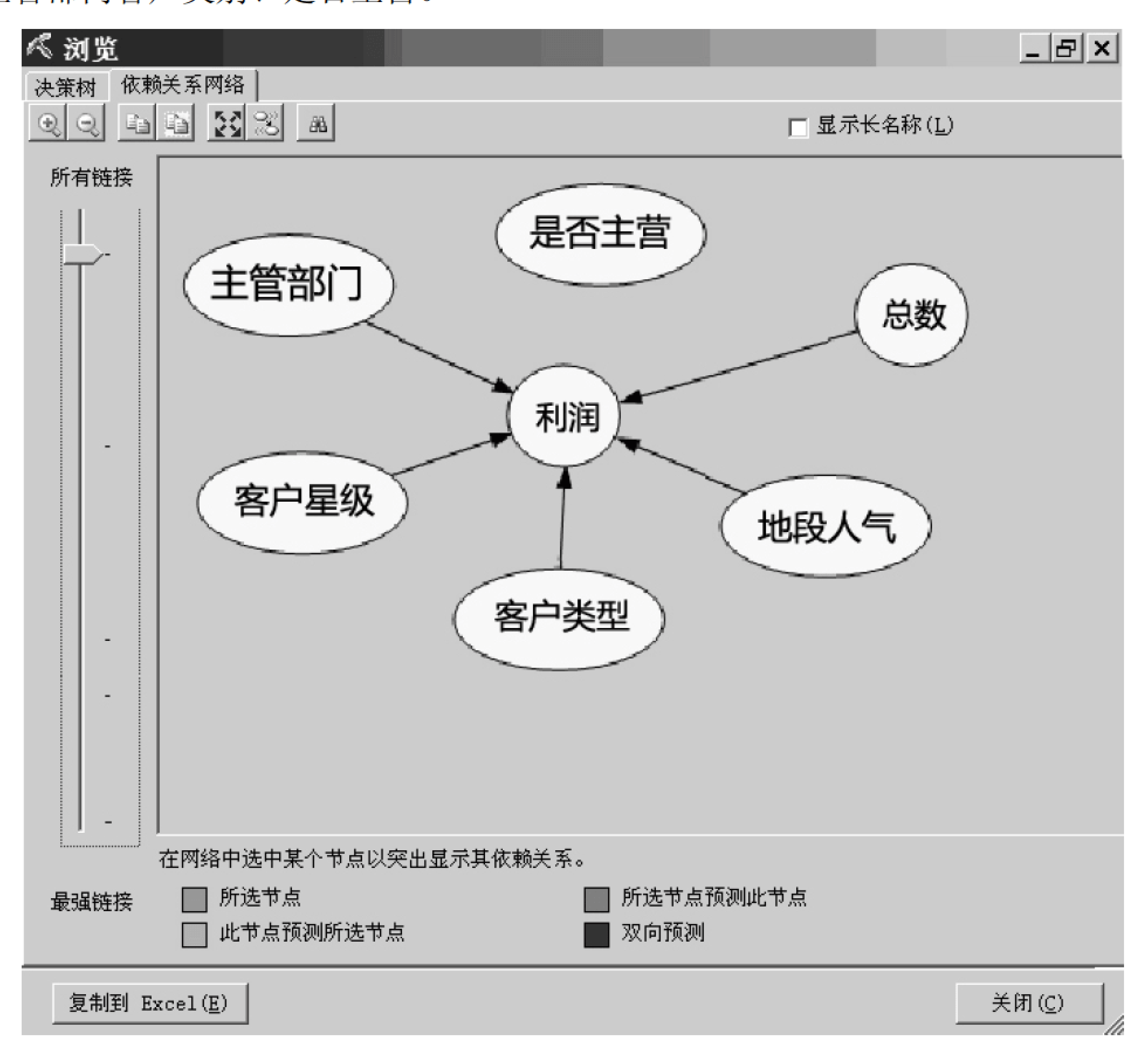


图 26-23　决策树 1 自变量与因变量利润的相依性

26.4.5　Logistic 回归

以客户星级为因变量，客户资料中的其他变量为自变量，构建 Logistic 回归模型 log，如表 26-11 所示。找出其他变量与客户星级的关系。通过构建 Logistic 回归模型，可以反映出不同星级客户在其他属性上的区别，如图 26-24、图 26-25、图 26-26 所示。

表 26-11　log 模型的变量列表

输入自变量名	含　　义	类　　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量

续表

输入自变量名	含　义	类　型
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量
利润	某商户月卷烟零售利润	连续变量



图 26-24　log 模型中的 4 星客户与 3 星客户的区别



图 26-25　log 模型中的 3 星客户与 2 星客户的区别



图 26-26　log 模型中的 4 星客户与 2 星客户的区别

1．模型结果

从图 26-24、图 26-25、图 26-26 可以看到模型的构建结果，比较三个星级的客户在其他属性上有什么不同。例如，当商户的从业人数为 25 或者 12 时，客户等级为 4 星的可能性远大于 3 星。当主管部门为省土产日杂公司时，客户等级为 2 星的可能性远大于 3 星、4 星。

2．挖掘精确度图表

如果对某种客户星级感兴趣，可以画出该 log 模型对此品牌的准确性图表，观察其预测效能。例如图 26-27，如果关心对“3 星”的客户的预测效能，可以画出其准确性图表。发现 log 模型在预测“3 星”的客户时，显著好于随机猜测，与理想模型相比，差距不大。

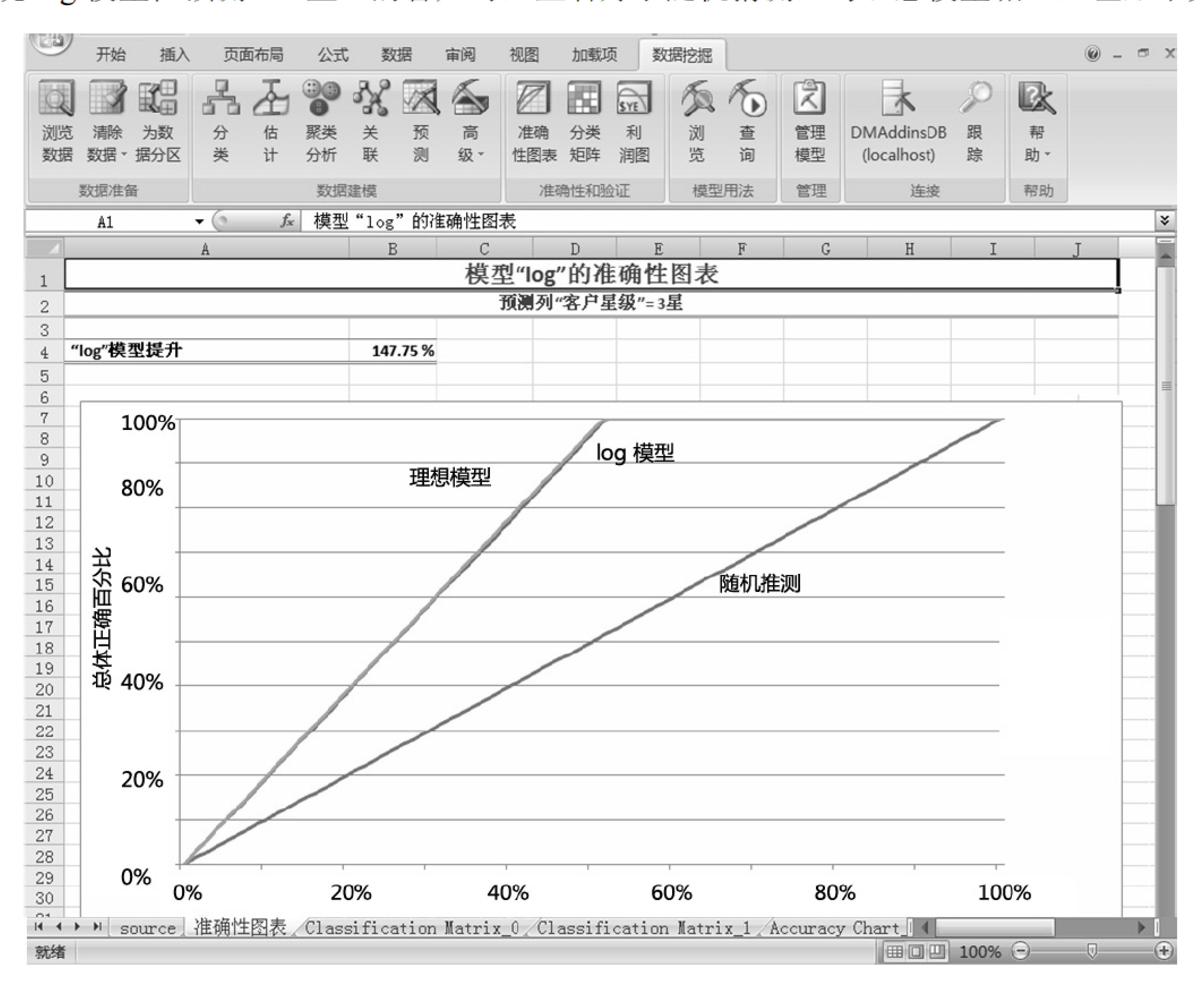


图 26-27　log 模型中的 3 星客户的预测准确性图表

表 26-12 为 log 模型的分类矩阵，由分类矩阵中亦可看出该模型能够比较准确地找出 3 星和 4 星的客户，但是会把大部分 2 星客户错估为 3 星客户，如图 26-28 所示。

表 26-12　log 模型的分类矩阵　%

	2 星（实际）	3 星（实际）	4 星（实际）
2 星	37.50	0.13	0.00
3 星	62.50	99.87	0.00
4 星	0.00	0.00	100.00

续表

	2 星（实际）	3 星（实际）	4 星（实际）
正确	37.50	99.87	100.00
分类错误	62.50	0.13	0.00

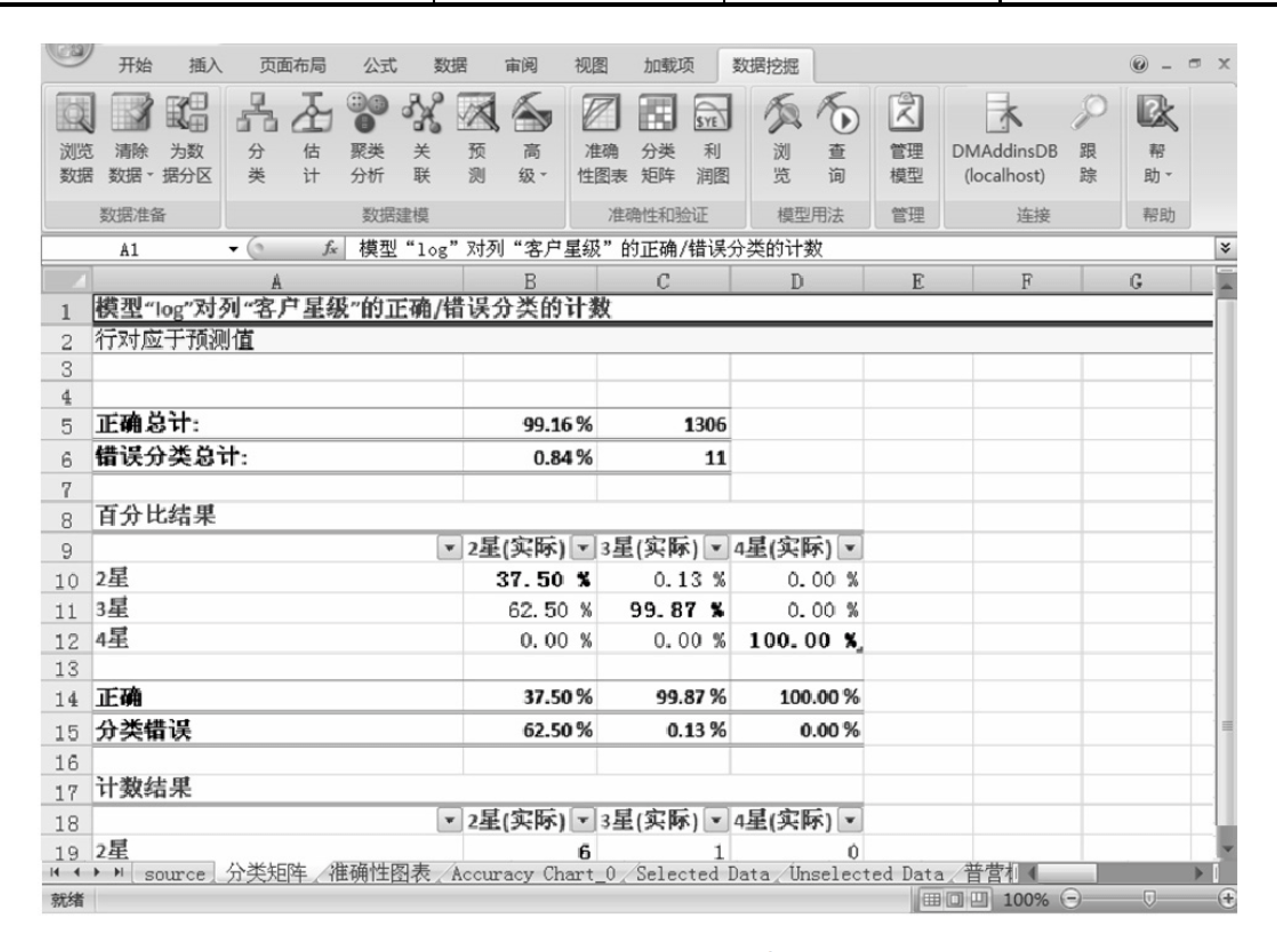


图 26-28　log 模型预测分类矩阵

26.4.6　关联分析

以客户星级为因变量，其余的客户资料变量为自变量，建立关联模型 association，如表 26-13 所示。借此找出哪种自变量和某种客户星级最有关系。

表 26-13　association 模型的变量列表

输入自变量名	含　　义	类　　型
出样能力	柜台陈列卷烟样品的数量	整型变量
出样形式	陈列卷烟样品的形式	分类变量
从业人数	该商户的员工人数	整型变量
地段人气	店面所处地段的繁华程度	分类变量
订货类型	电话或网络等订货方式	分类变量
结算方式	付货款的方式	分类变量

续表

输入自变量名	含　义	类　型
客户类别	一种烟草管理部门的评价	分类变量
客户星级	一种烟草管理部门的评价	分类变量
入网日期	零售户加入销售网络的日期	日期变量
是否主营	主营卷烟还是兼营	分类变量
文化素质	店主（法人代表）的学历	分类变量
许可证种类	持有何种零售许可证	分类变量
营业面积	店面的营业面积	连续变量
主管部门	所在地的烟草零售主管部门	分类变量
总数	某商户月销售总数	整型变量
利润	某商户月卷烟零售利润	连续变量

构建模型完毕，可以用依赖关系网络来反映因变量的不同取值和自变量的不同取值之间的关系。图 26-29 反映客户星级为 3 星的关联状况。此种客户星级和兼营这种经营形式关联最强，此外，还和出样能力小于 18，月销售量小于 6 511 条，利润小于 2 652.49 元等条件有关系。

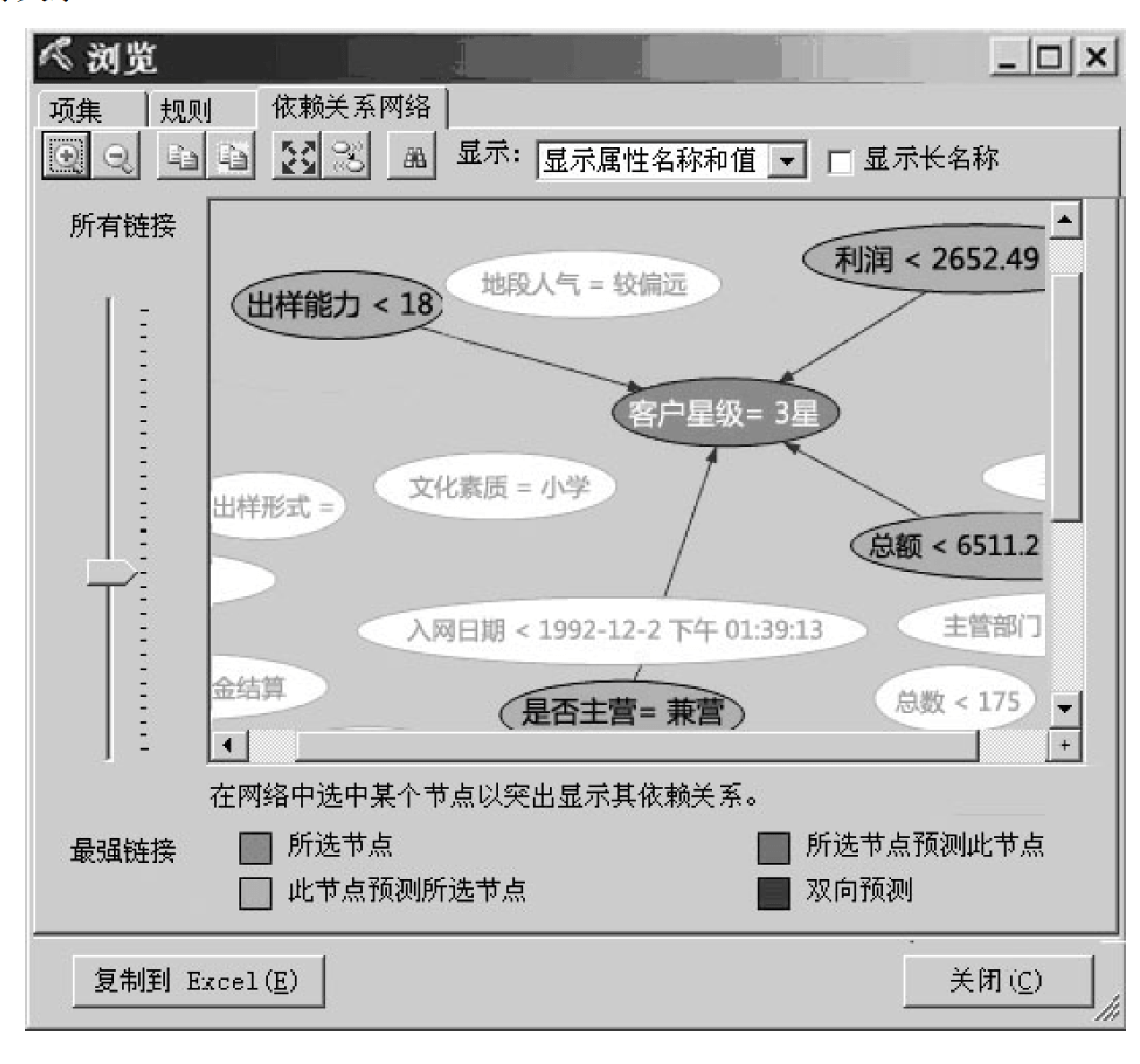


图 26-29　association 模型依赖关系网络——客户星级为 3 星

图 26-30 反映客户星级为 4 星的关联状况。此种客户星级和店主文化素质是高中关联

最强，此外，还和月销售额、月销售量、利润等很多属性有关系。

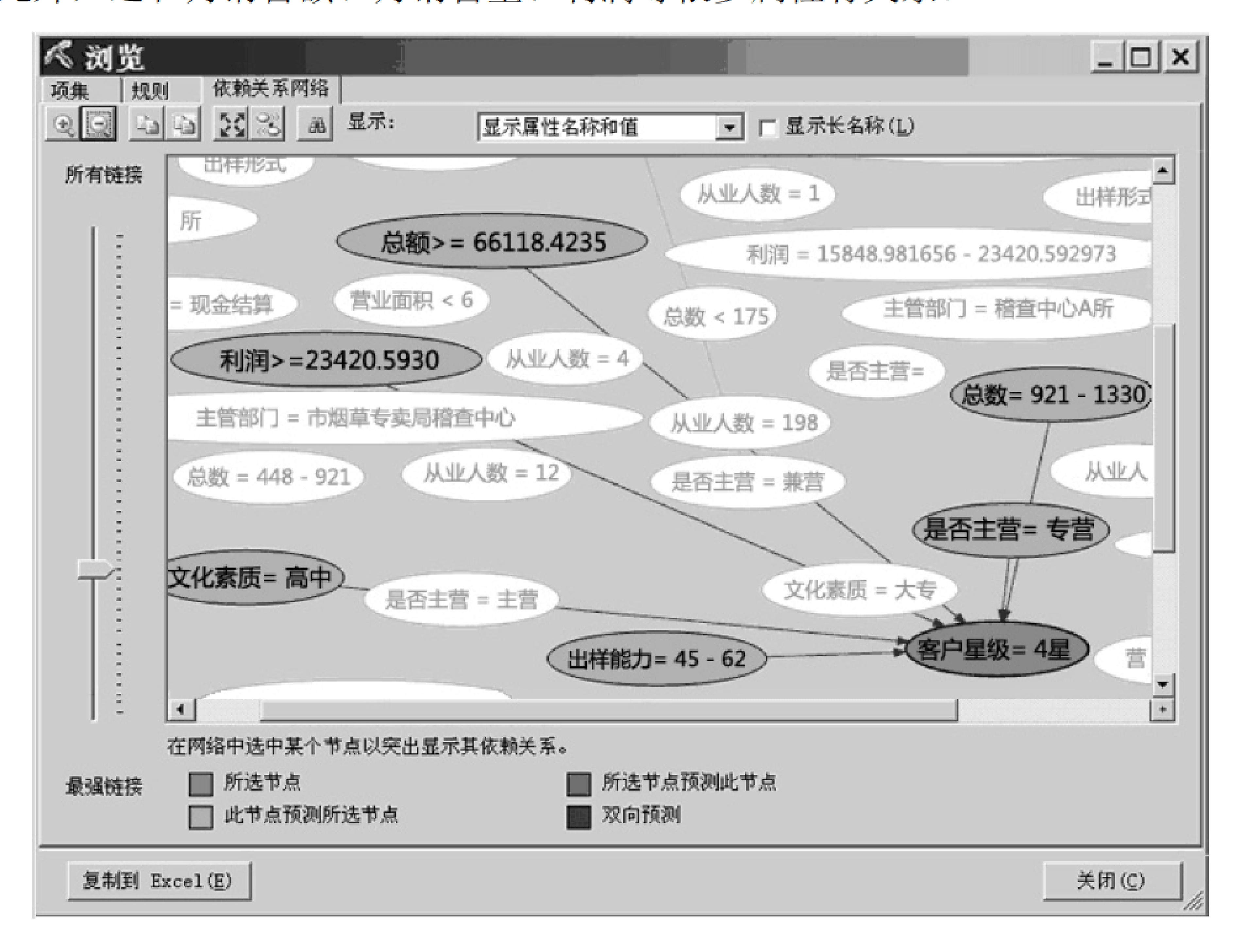


图 26-30　association 模型依赖关系网络——客户星级为 4 星

图 26-31 反映客户星级为 2 星的关联状况。此种客户星级和主管部门为省土产日杂公司、从业人数为 25 这两个条件有明显的关联关系。

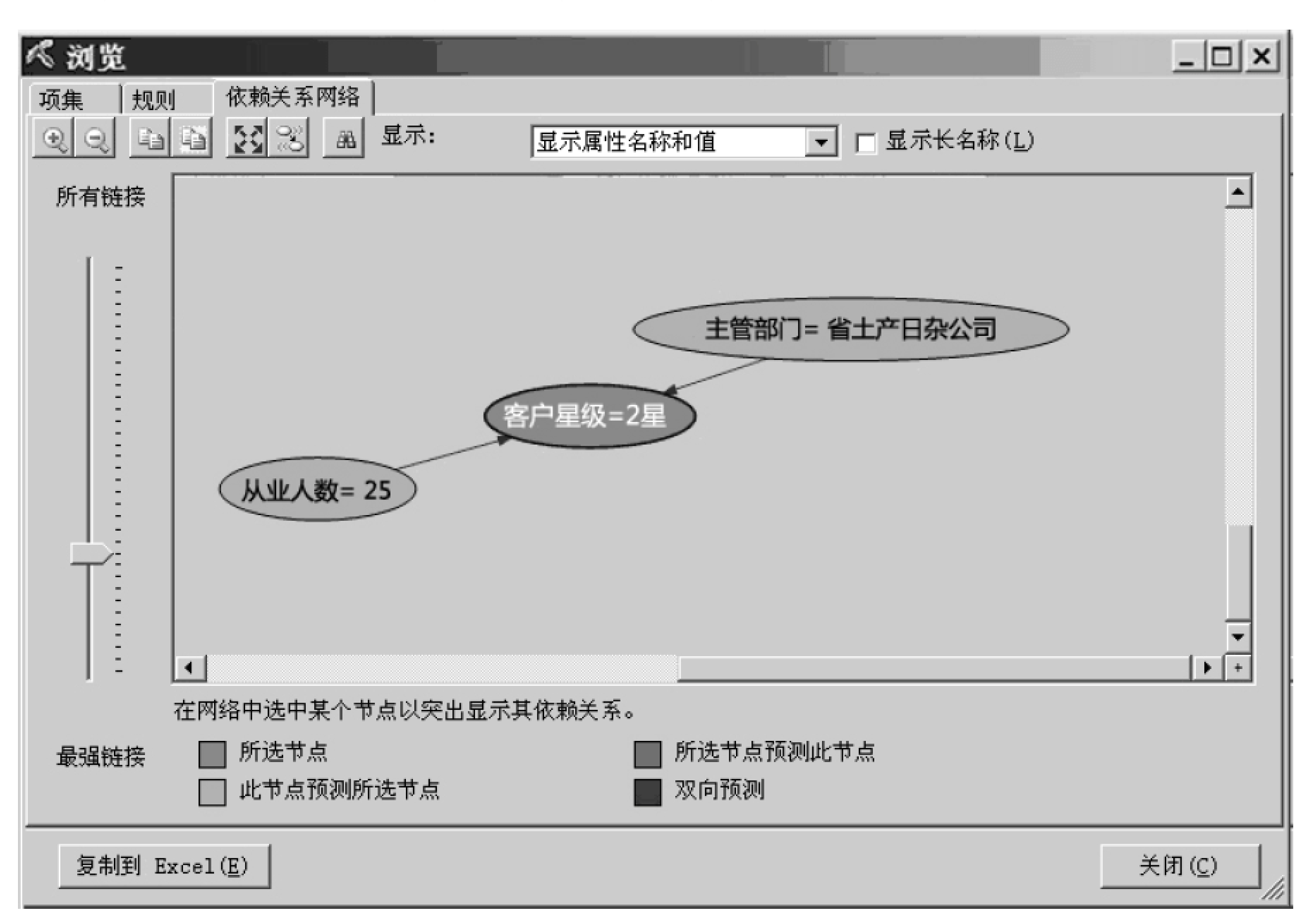
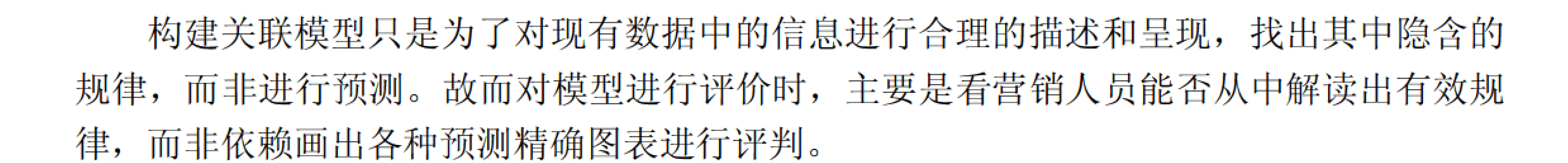


图 26-31　association 模型依赖关系网络——客户星级为 2 星



构建关联模型只是为了对现有数据中的信息进行合理的描述和呈现，找出其中隐含的规律，而非进行预测。故而对模型进行评价时，主要是看营销人员能否从中解读出有效规律，而非依赖画出各种预测精确图表进行评判。

26.5　结论

依据以上分析与建模，有以下的结论整理：

（1）决策树模型，反映某零售户所购买卷烟的等级与卷烟的利润率有最强的关联性，其次为月销售总数及主管部门。

（2）贝叶斯概率，在会选购 11mg 滤盖红双喜的零售户的属性中，以“店面在繁华交通要道，主管部门为稽查中心晋源所，月销售总数在 201 条以上，客户星级为 4 星”的居多。

（3）决策树中，显示月总销售数与利润最有关联性，其次依序为地段人气、客户星级、主管部门客户类别、是否主营。

（4）Logistic 回归中，由于零售户等级多为 3 星和 4 星，因此比较 3 星与 4 星的差异。主管部门为省土产日杂公司，从业人数为 12 或 25 人，客户类别为 A 类的零售户为 4 星等级的可能性较大；而从业人数为 7 或 198，主管部门为烟草实业公司或粮油总公司的零售户，偏向 4 星的可能性大。

（5）关联分析中，客户星级为 3 星的零售户有很大可能性：出样能力小于 18，月销售量小于 6 511 条，利润小于 2 652.49 元。客户星级为 2 星的零售户，主管部门为省土产日杂公司和从业人数为 25 的可能性较大；客户星级为 4 星的零售户，店主文化素质是高中，月销售总额大于 66 118 元，销售总数在 921～1 330 条之间，利润大于 23 420 元，出样能力在 45～62 之间，经营类型为专营的可能性最大。